DOCUMENT RESUME

ED 256 082 EC 170 489

AUTHOR	Scruggs, Thomas E.
TITLE	The Administration and Interpretation of Standardized Achievement Tests with Learning Disabled and Behaviorally Disordered Elementary School Children. Final Report.
INSTITUTION	Utah Univ., Salt Lake City.
SPONS AGENCY	Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE	2 Jul 84
NOTE	172p.; Developed at the Developmental Center for the Handicapped. For the test taking skills training materials, see EC 170 490.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Achievement Tests; Attention Control; *Behavior Disorders; Elementary Education; *Learning Disabilities; *Student Attitudes; Test Anxiety; Test Coaching; *Test Wiseness

ABSTRACT

Several experiments were carried out to determine: (1) whether learning disabled (LD) and behaviorally disordered (BD) students exhibit deficiencies with respect to appropriate test-taking strategies and (2) if so, whether these strategies could be successfully trained. In the test-training evaluation, 92 LD or BD elementary-age students representing grades 2, 3, and 4 were randomly assigned to treatment or control conditions. Treatment subjects received eight training sessions on test-taking skills, with particular regard to the Stanford Achievement Test. All treatment students scored significantly higher on a test of test-taking skills. In addition, third and fourth grade LD and BD students scored significantly higher on the word Study Skills subtest and exhibited descriptive increases with respect to other subtests. Second grade students were apparently unaffected by the training procedure. A similar test-training package applied to intact third grade classrooms of mostly nonhandicapped students indicated that these materials were effective in improving student attitudes toward the test-taking experience. The document begins with a project overview and contains the following project manuscripts: "Improving the Test-Taking Skills of LD and BD Elementary Students" (C. Taylor and T. Scruggs); "An Analysis of Children's Strategy Use on Reading Achievement Tests" (T. Scruggs, K. Bennion, and S. Lifson); "Developmental Aspects of Test-Wiseness for Absurd Options: Elementary School Children" (T. Scruggs); "Format Changes in Reading Achievement Tests: Implications for Teachers" (K. Bennion, S. Lifson, and T. Scruggs); "Passage Independence in Reading Achievement Tests: A Follow-Up" (S. Lifson et al); "Spontaneously Employed Test-Taking Skills of Learning Disabled Students on Reading Achievement Tests" (T. Scruggs et al); "Spontaneously Employed Test-Taking Strategies of High and Low Comprehending Elementary School Children" (T. Scruggs et al); "Teaching Test-Taking Skills to Elementary Grade Students: A Meta-Analysis" (T. Scruggs et al); "The Effects of Training in Test-Taking Skills on Test Performance, Attitudes, and On-Task Behavior of Elementary School Children" (T. Scruggs et al); and "Teaching Test-Taking Skills to Learning Disabled and Behaviorally Disordered Children" (T. Scruggs). (CL)

ED256082

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The Administration and Interpretation of Standardized Achievement Tests with Learning Disabled and Behaviorally Disordered Elementary School Children

Final Report Submitted
to
Special Education Programs
(CFDA 84.023C)
U.S. Department of Education

July 2, 1984

Dr. Thomas E. Scruggs (801) 750-1224
Developmental Center for the Handicapped
Utah State University

EC170489

Abstract

Several experiments were carried out over the course of a 12-month period to determine whether: (a) learning disabled (LD) and behaviorally disordered (BD) students exhibit deficiencies with respect to appropriate test-taking strategies, and, if so, (b) whether these strategies could be successfully trained. Preliminary investigations indicated that mildly handicapped students do exhibit deficiencies in the area of test-taking strategies. These deficiencies include attention to inappropriate distractors, failure to successfully employ prior knowledge and deductive reasoning strategies, and failure to identify correctly specific types of questions which call for different strategies. In the test-training evaluation, approximately 100 LD and BD elementary-age students representing grades 2, 3, and 4 were randomly assigned to treatment and control conditions. Treatment subjects received eight training sessions on test-taking skills with particular regard to the Stanford Achievement Test. All students scored significantly higher on a test of test-taking skills. In addition, third and fourth grade LD and BD students scored significantly higher on the Word Study Skills subtest and exhibited descriptive increases over experimental group with respect to other subtests. Second grade students were apparently unaffected by the training procedure. In addition, a similar test-training package applied to intact third grade classrooms of

mostly nonhandicapped students indicated that these materials were successful in improving student attitudes toward the test-taking experience.

PROJECT OVERVIEW

The primary objective of this project was to determine whether scores on standardized achievement tests could be improved through a combination of reinforcement, practice, and training of "test-taking skills"; that is, those skills which refer to understanding of the most efficient means to take a test, rather than knowledge of the content area. Such training, if successful, would likely improve the validity of resulting test scores in that a potential source of error, i.e., difficulty with format, testing conditions, etc. would be eliminated. In addition to the major objectives, several smaller investigations were planned and carried out, the ultimate objective of which was to determine whether, in fact, students in special education placement exhibited specific deficiencies on select aspects of test-taking. In addition, another test-training investigation was carried out on intact third grade classes in order to determine whether such a training package was appropriate to whole class administration and whether such training produced any change in on-task behavior or attitudes toward the test-taking experience. Approximately 15% of this population was classified as learning disabled or behaviorally disordered.

Preliminary Investigations

In general, the project has proceeded in accordance with the planned schedule of activities in the proposal. However, when the proposal was prepared, it was assumed that materials development would not be necessary as materials had been developed from a prior project and were at that time being validated. Since this project was funded, however, it has been determined that those materials as implemented were not effective in increasing the performance of students in regular education classes on standardized achievement tests. It was, therefore, thought necessary to initiate a series of studies to evaluate what specific skills lower functioning students may lack with respect to test taking, and to develop a new set of materials which might more specifically address these needs. Accomplishments are described below by each task.

1. Assessment of spontaneously employed test-taking strategies (July-December, 1983). A shorter version of the Stanford Achievement Test, Reading subtests, questionnaire form and follow-along sheet, were developed in order to evaluate the skills students spontaneously employed in test-taking situations. These materials were utilized in several studies to acquire this information. Students were selected from two remedial and one original program from each of grades 1 through 7. Students were individually administered selected subtests of the Stanford Achievement Test. They were asked for their level of confidence for each answer and the strategies they had chosen for answering the questions. It was determined that a complete hierarchy of

strategies existed with respect to answering test questions beyond simply knowing or not knowing the answer, and that these strategies resulted in differential levels of performance on the part of the students. It was generally seen that younger students and academically lower functioning students tended to produce lower-level strategies than higher functioning and older students. This investigation is described in detail in the manuscript in the appendix entitled, "An Analysis of Children's Strategy Use on Reading Achievement Tests". This manuscript has been accepted for publication by Elementary School Journal. Additional evaluation of the data from this investigation indicated the existence of a developmental trend through the elementary grades in the use of elimination strategies on ambiguous multiple choice items. That is, as children got older, they became more proficient with respect to their spontaneous ability to eliminate inappropriate or obviously incorrect alternatives. These results have also been described in detail in the manuscript entitled, "Developmental Aspects of Test-Wiseness for Absurd Options: Elementary School Children". This manuscript has been submitted for publication.

A test of "passage independence" of reading comprehension test items on the Stanford Achievement Test was developed by administering items from the Reading Comprehension subtest of the SAT to college undergraduates. The purpose of this investigation was to determine what proportion of these test items were potentially answerable by employing prior knowledge or deductive reasoning skills. It was determined that college undergraduates were able to answer nearly 80% of these questions on the average, with many students answering them all correctly. This article is given in the appendix under the title, "Passage Independence in Reading Achievement Tests: A Follow-Up," and has been published in the journal, Perceptual and Motor Skills.

Two follow-up investigations were intended to examine more precisely the nature of test-taking strategies employed by learning disabled students, specifically as compared with the strategies employed by their non-disabled counterparts. In one investigation, LD and non-LD students were administered items from the Stanford Achievement Test, Reading Comprehension Subtest, with the actual reading passages deleted from the test. Students were told to simply answer the questions the best that they could. In the second experiment, all items were read to both groups of students in order to control for general reading ability. In both experiments, students not classified as learning disabled scored significantly higher on this test of "passage independent" test items than did their learning disabled counterparts. These results indicated (a) that learning disabled students may differ with respect to spontaneous test-taking strategies, such as use of prior knowledge and deductive reasoning skills, and (b) raise the issue of what such test items are actually measuring, since they could be so easily answered without having read the corresponding passage. This investigation has been written in manuscript form, and is in the appendix under the title, "Are Learning Disabled Students Test-Wise: An Inquiry into Reading Comprehension Test Items", and has been submitted for publication.

In a second investigation, learning disabled and non-learning disabled students were directly questioned with respect to strategies they employed on reading comprehension test items and letter sounds test items. In this investigation, it was found that learning disabled students did not differ from their non-disabled peers with respect to answering recall comprehension questions, with ability to read controlled. However, learning disabled students were less likely to employ appropriate strategies to answer inferential questions, and reported inappropriately high levels of confidence in their responses. In addition, when they did report using appropriate strategies, they were much less likely to employ them successfully. This project is described in detail in the manuscript, "Spontaneously Employed Test-Taking Skills With Learning Disabled Students on Reading Achievement Tests." This manuscript has also been submitted for publication and was presented at the annual meeting of the American Educational Research Association in New Orleans in April.

In an investigation which has not yet been reported, it was determined that a sample of elementary-age behaviorally disordered students scored significantly lower ($\underline{t}(35) = 2.59$, $\underline{p} < .01$) than their nonhandicapped counterparts with respect to reported attitudes towards tests and the test-taking situation. These investigations, taken together, provided valuable information regarding the most optimal training package to be developed for use with this population.

An evaluation of all major achievement tests was also made in order to determine whether tests were similar or different with respect to format demands on the test taker. In this investigation, all levels of six major achievement tests were evaluated for number of format changes per minute throughout the reading achievement test subtest. It was determined that achievement tests varied widely with respect to format demands, with most format changes occurring in the primary grades. These results are documented in the manuscript, "Format Changes in Reading Achievement Tests: Implications for Learning Disabled Students," which can be found in the appendix and has been submitted for publication.

In order to evaluate appropriately all previous attempts to train test-taking skills in the elementary grades, a meta-analysis was completed of all available studies in this area. It was determined that although the general effect of training was positive, differences in favor of training groups did not seem to become substantial unless training was relatively extensive. In addition, this meta-analysis revealed that low SES children and primary grade children were more likely to benefit from extended training hours. This seems to underline the importance in the present project of implementing a package of a higher level of intensity. The detailed results of this meta-analysis are given in the appendix under the title, "Improving Achievement Test Scores in the Elementary Grades by Coaching: A Meta-Analysis." This manuscript has also been submitted for publication.

Finally, during the first part of the project, the scope of the proposed research was described and published by Exceptional Children in the fall of last year and is given in the appendix under the title, "Research in Progress: Improving the Test-Taking Skills of Learning Disabled and Behaviorally Disordered Elementary School Children." In addition, during the fall, preliminary findings were reported at the seventh annual conference of Severe Behavior Disorders of Children and Youth in Tempe, Arizona in a presentation entitled, "Training Behaviorally Disordered Children to Take Tests."

It was the intention of all of the above investigations to evaluate both tests and test-taking strategies of mildly handicapped students in order to determine the most likely strategies for intervention and the form that intervention should take. In all, it was determined that mildly handicapped students do differ from their nonhandicapped peers with respect to use of appropriate strategies on standardized achievement tests. It was also determined that these strategy deficits included use of prior knowledge, use of deductive reasoning skills, attention to appropriate distractors, and selection of strategies appropriate to correctly answering different types of items.

2. Development and revision of training materials (September-February, 1983-1984). Based upon results of the above investigation and careful evaluation of the Stanford Achievement Test, materials were developed which were intended to teach to second, third, and fourth grade children in special education placements skills appropriate to the successful taking of the Stanford Achievement Test. These materials included eight scripted lessons and a student workbook of exercises on subtests meant to be very similar to those used on the Stanford Achievement Test. These materials were intended to teach both general test-taking strategies, such as efficient time usage, as well as specific lessons meant to increase understanding of the particular test demands of the individual reading subtest of the Stanford Achievement Test. These materials are included with this report and are entitled "Super Score."

Following the preliminary development of materials, they were pilot tested in November on two groups of second grade children with learning and behavioral disorders. On the basis of this pilot investigation, several revisions were made in the materials. Specifically, some of the lessons proved to be too long for the most effective implementation with this project, and some instructions were judged to be ambiguous. In addition, a pre and posttest measure which was developed for use with this population was also judged to be inadequate to effectively assess progress made on these materials.

On the basis of the initial pilot investigation, the materials were revised and expanded to include second to fourth grades, and were then implemented in a larger field test involving 24 students in special education placements in second, third, and fourth grades. Students were randomly assigned to treatment and control groups at each of the three grade levels, and the lessons were administered to the treatment groups. Students in the experimental group were seen to score higher than students in the control group on a shortened version of the Stanford Achievement Test, Reading Subtest.

These findings were not conclusive due to the small number of subjects employed in each grade, and the fact that different forms of the Stanford Achievement Test appeared to have been differentially difficult for different grade levels, the result being a differential level of difficulty on the posttest measure. Although statistical significance was not found, it was determined that students in the experimental group had scored .48 standard deviation units higher than students in the control group on the Reading Achievement Word Study Skills and Reading Comprehension subtests. This effect size, had it been a reliable one and occurred in the primary grades on an actual test administration, would have been equivalent to a four- or five-month gain score for students who had received the training. In addition, analysis of pre- and posttests of test-taking skills indicated that the materials had in fact been effective in training these particular skills.

Some final revisions were made of the training materials on the basis of the second field test, and materials were finally prepared for spring implementation immediately prior to district-wide standardized test administration. While final revisions were being made, individual schools were contacted to be involved in a larger experimental study intended to validate these materials. For this study, approximately 110 students enrolled in special education classes in grades 2, 3, and 4 in two different large elementary schools were selected and randomly assigned to treatment and control conditions. Four persons, including the principal investigator, took part in the two-week training period which was administered at the end of March. This training was administered in eight 20- to 30-minute sessions given from Monday to Thursday for each of two weeks immediately prior to district-wide test administration. At the same time, materials were developed intended to increase test-taking skills on the Comprehensive Test of Basic Skills and were administered in the school districts adjacent to Utah State University. This training package was implemented in local third grade classes in order to determine (a) whether these procedures were appropriate for whole-class administration, (b) whether the materials developed for the Stanford Achievement Test could be easily adapted to other tests, and (c) whether such training could be seen to have an impact upon test scores, attitudes, and time on-task during test administration.

The results of the training on the Comprehensive Test of Basic Skills in the local third grade classes indicated that students' attitudes had, in fact, qualitatively improved as a result of the test training. It was suggested that the test training had resulted in a more normal distribution of attitudes after the end of the three days of testing and implied that the training had made the test-taking experience itself less traumatic on the part of third grade regular classroom students (including 15% mildly handicapped students). Time on-task during directions and during the test-taking experience itself did not seem to be affected by the training package. In addition, the training was seen to significantly increase the scores of students in the lower half of the class on the Word Attack subtest of the reading test. Analysis of the top half, or the group as a whole, was not possible due to the presence of strong ceiling effects in both experimental and control groups. This investigation has been written in manuscript form and is given in the appendix under the title, "The Effects of Training on the Standardized Test Performance, On-Task Behavior, and Attitudes of Third Grade Children." This manuscript has been submitted for publication.

Results of the training package with second, third, and fourth special education students also indicated that the training was successful in improving scores on standardized achievement tests. Although only descriptive differences were seen in some subtests, the training package significantly improved the performance of the experimental students over control students in the Word Study Skills subtest. This improvement was judged to be approximately equivalent to a three- to four-month increase in equivalent grade level. The fact that improvement in the Word Study Skills subtest was observed was considered to be due to the fact that this particular subtest involved many smaller subtests, several format changes, and potentially confusing directions for which the training package was thought to have been particularly helpful. Descriptive differences were seen in other subtests of the SAT but, not being statistically significant, it is not possible to determine whether they were a result of the training or simply sampling error. Evaluation of scores of the second grade students indicated that they apparently had not benefited from the training package. However, the differentially small number of subjects in the second grade sample, attrition suffered during the training, and the fact that these two groups were in retrospect found to have differed with respect to the previous year's testing, obscure clear interpretation of this data. It may be, for example, that second grade LD and BD students have insufficient reading and other academic skills to enable them to benefit from this training package, or it could be that these students had in fact benefited but that due to sampling and attrition problems, these benefits were not observed. This entire investigation has been described in detail and is given in the appendix under the title, "Training Test-Taking Skills to Learning Disabled and Behaviorally Disordered Students," which has been submitted for publication.

Conclusions

The major findings of the year's research suggest that: (a) mildly handicapped students differ from their nonhandicapped peers with respect to spontaneously employed test-taking strategies and attitudes toward the test-taking situation, and (b) that these test-taking skills and attitudes can be signficantly improved by training. These findings indicate that for children classified as learning disabled or behaviorally disordered, achievement test scores often may not be as accurate a measure of actual academic performance as is possible. It also seems to indicate that training to increase test-taking skills and attitudes towards tests may significantly increase the individual handicapped student's functioning on these tests.

A case can be made that norm-referenced tests are not solely relied upon in making placement decisions, and that in fact other individually administered tests are better indicators of specific skill deficits with teaching implications. It is true, however, that these students deserve to be taught basic skills that they may lack in any particular area, including taking standardized group administered achievement tests, and that if their poor performance can be improved at all, this seems to indicate that substantial error has been reduced from the tests. Any such improvement then is judged by the present project personnel to be worthwhile.

Several questions, however, remain to be investigated by the present project. First, whether or not this type of training is likely to result in increased scores on math subtests is completely unknown and, in fact, cannot be determined on the basis of the present investigation. In addition, the extent to which secondary-age learning disabled and behaviorally disordered children are deficient in test-taking skills and attitudes and to what extent these may be trainable also cannot be concluded from the present investigation. It is the purpose of the project during the second year to investigate test-taking deficiencies on math subtests and corresponding potential for training, and the third year, to evaluate test-taking characteristics of secondary-level learning disabled and behaviorally disordered students. It is the hope of this project that by the third year of funding, general conclusions can be made with respect to mildly handicapped learners of all ages and several different types of achievement tests. It is hoped that this information will be of benefit to many special educators, and particularly students in special education classes, throughout the country.

APPENDIX: Project Manuscripts

BEST COPY AVAILABLE

RESEARCH IN PROGRESS

Charles C. Cleland
Department Editor

Improving the Test-Taking Skills of LD and BD Elementary Students

Principal Investigators: Cie Taylor and Thomas Scruggs, Exceptional Child Center, Utah State University.

Purpose/Objectives: The purpose of this investigation is to determine whether reinforcement techniques and direct training in test-taking skills can increase the validity of test scores for learning disabled (LD) and behaviorally disordered (BD) students. To determine the degree to which LD and BD students exhibit inappropriate (inefficient) test-taking skills, students are observed and interviewed while taking standardized tests. Based on those observational data, procedures and training packages will be designed to increase student performance on standardized achievement tests. If the procedures and training are effective, educational decisions, which are frequently based in part on the results of standardized achievement tests, will be more valid because problems in areas such as test-taking skills, student motivation, and confusion due to testing format will be reduced or eliminated.

Subjects: Subjects are 100 elementary students enrolled in 12 resource rooms and self-contained classrooms for children with learning disabilities and behavioral disorders.

Methods: LD and BD children matched on age, handicap, and standardized achievement test score will be randomly assigned to experimental and control groups. Students in the experimental group will receive materials and procedures designed to improve the ability of handicapped students to take tests. Experimental and control groups will be compared statistically on several measures, including attitudes toward test-taking, student and teacher behavior during test administration, and actual performance on standardized tests of reading achievement. In following years, materials will be developed and implemented for mathematics achievement tests and test-taking skills for secondary-age handicapped students.

Results to Date: Preliminary findings indicate that many LD and BD children, as well as low achieving nonhandicapped students, do not spontaneously exhibit efficient test-taking behaviors. Specifically, handicapped children have been seen to exhibit difficulties with item format and distractors more typical of naive test takers.

Commencement and Estimated Completion Dates: This investigation began July 1, 1983 and is expected to continue for three years.

Funding: Funding for this investigation has been provided by a grant from the U.S. Department of Education, Research in Education of the Handicapped.

Publications/Products Available: Preliminary materials for improving test-taking skills, piloted on nonhandicapped second-grade students, have been developed and will be revised for use with handicapped children during the coming year. Manuscripts documenting the investigation will be completed and submitted for publication during the second half of the academic year. Please write the authors for further information.

"Research in Progress" is a forum for reporting ongoing research in the field of special education that has not yet been published. Investigators wishing to report studies in progress are invited to submit a brief synopsis of their efforts to the column editor, Charles C. Cleland, 3427 Monte Vista, Austin TX 78731. Reports are to be submitted in triplicate and should follow the format shown above, with a maximum length of 500 words.

An Analysis of Children's Strategy Use on Reading Achievement Tests[1]

Thomas E. Scruggs
Karla Bennion
Steve Lifson
Exceptional Child Center
Utah State University

Running head: Children's Strategy Use

An Analysis of Children's Strategy Use on Reading Achievement Tests

Much of what constitutes reading instruction in today's public schools reflects students' scores on standardized achievement tests. Test performance may influence later assignment into reading groups or classrooms, or remedial or special education programs. Although norm-referenced reading tests have been criticized as being insensitive to specific skill deficits and inadequate as complete diagnostic measures (Howell, 1979), most reading tests have nonetheless been seen to be highly reliable and valid (Spache, 1976). For better or worse, standardized reading tests are very much a part of education today and will most likely continue to be used in the future.

If important decisions are to be based on the results of standardized reading tests, student scores should provide the best possible estimate of reading performance. Unfortunately, the results of past research have indicated that student reading test performance can be influenced by factors other than knowledge of test content (e.g., Taylor & White, 1982). One of these factors, test-wiseness (TW), was first described in detail in 1965 by Millman, Bishop, and Ebel as "a subject's capacity to utilize the characteristics and formats of the test and/or the test-taking situation to receive a high score" (p. 707). Millman et al. developed an outline of test-wiseness principles, which included time using strategies, error avoidance strategies, guessing strategies, and deductive reasoning strategies. Slackter, Koehler, and Hampton (1970) presented information which suggests that TW has a developmental component. That is, students may become more "test-wise" as they grow older. Generally, researchers have inferred extent of TW on the basis of tests specifically constructed for this purpose.

Recently, students themselves were questioned about strategies they use to answer test questions. Haney and Scott (1980) administered a number of achievement tests to 11 students, then questioned each student the following day concerning the manner in which they attempted to answer each item. These researchers developed a complex model with which responses to interviewer questions were classified into 46 separate categories. Most of these categories included the use of some specific strategies such as guessing, elimination of alternatives, or "reasoning." Their results indicated that children use a wide range of strategies in answering test questions and that often the child's perception of item content bears little resemblance to the intention of the author of the test. Haney and Scott concluded that considerable "ambiguity" exists in standardized test questions and that it exists to a greater extent in science and social studies areas, and to a lesser extent in reading areas.

The work of Haney and Scott contributed significantly to our knowledge of the nature of ambiguous test items. The focus of their study, however, was on test construction, with implications concerning the reduction of test item ambiguity. Although classroom teachers may use the results of Haney and Scott to improve their own tests, published standardized tests cannot be altered by teachers. A question which remains concerns the extent to which students employ "test-taking" strategies when faced with difficult or ambiguous items. Do students spontaneously use such strategies (that is, without being trained)? If so, which strategies (if any) are effective in obtaining correct answers? No previous research has been located to answer these questions.

To address those questions in the present study, the reading test performance of elementary school children was examined. Specifically, two areas were investigated: (a) the strategies spontaneously employed by students to answer reading test items, and (b) the relative effectiveness of these strategies in increasing reading test scores.

Procedure

A sample reading test based upon items from the Stanford Achievement Test (SAT) was developed and piloted on five students to evaluate whether the length was appropriate and to establish reliable scoring conventions. This sample test included items from the Word Reading, Reading Comprehension, Word Study Skills, and Vocabulary subtests. After revisions had been made, it was administered to 31 elementary age Caucasian students (15 girls, 16 boys) attending summer classes in a western rural area. Students were selected from both remedial and "enrichment" classes so that a range of abilities was represented. Twenty students were seen to read at or above grade level; 11 were seen to read below their grade level as assessed by the Woodcock Reading Achievement Test. Most students (20) were second or third graders, but students were also selected from grades 1 (2), 4 (2), 5 (5), and 6 (2).

All students were seen individually by one of four examiners. One examiner interviewed 18 students, while the other three interviewed 4, 6, and 2 students. First, students were given the Woodcock Reading Achievement Test, Passage Comprehension subtest, in order to identify an approximate reading comprehension grade equivalent. Students were then given selections from the SAT taken from the level one year higher than their assessed grade level on the Woodcock subtest. In this manner, a similar difficulty level was provided for each student. Most students were able to answer correctly approximately two-thirds of the test questions.

Students were then told to read aloud each test question (as well as the reading passages in the reading comprehension subtest), and to read aloud whichever of the distractors they chose to read. They were neither encouraged nor discouraged from reading each distractor. As soon as students had answered a test question, they were asked to rate their level of confidence in their response: were they very sure, somewhat sure, or not sure the answer they had given was correct? After students had finished each subtest, they were asked to re-read the questions and tell the examiner why they had chosen the answer they did. The examiner recorded reading errors, confidence levels, attention to distractors, reference to reading passage, and reported strategies. Sessions were tape recorded to clarify any later ambiguity in scoring. Students spent 45-90 minutes in the session and answered 31-42 test questions. Some students received more questions than others because different levels of the SAT required different subtests and formats.

Results

Effectiveness of Strategies

We found that all strategy responses could be classified within a 10-level hierarchy which strongly predicted probability of correct responding. Proportion of correct responses were computed across subjects for each type of strategy used and are shown in Figure 1. These classifications were as

---

Insert Figure 1 about here

---

as follows: (a) skipped (student skipped the item), (b) misread a key word in question or distractors, (c) used faulty reasoning (example: one student reported, "this word must be the correct answer because it has a period after it"), (d) didn't follow directions, (e) guessed, (f) "seemed right" (student thought the answer was correct without being able to state an explicit reason), (g) used external information (example: "I know most people in fires die from breathing smoke because a fireman told me that"), (h) eliminated inappropriate alternatives, (i) referred to passage, and (j) clearly "knew" the answer (example: "I know that a pear is a kind of fruit"). The existence of these strategies indicated that a complete hierarchy of test-taking skills exists beyond simply knowing or not knowing the answers, and these strategies can be more or less effective on a standardized reading test. As seen in Figure 1, for example, when students skipped an answer, they got none correct; when they guessed, they got 37% correct; when they eliminated alternatives, they got 67% correct. Proportions of strategies employed are given in Table 1.

Insert Table 1 about here

We collapsed these strategies into five logical categories (skipping, procedural error, guessing strategy, deliberate strategy, and "knowing") and computed point biserial correlations for each subject. The median correlation between item score and reported strategy was .54 ($\underline{p} < .01$), a correlation of moderate strength which indicated that over 30% of the variance in test performance was held in common with the level of test-taking strategy employed.[2] No differential effects were seen by age, ability level, or examiner, although the sample was too small to conclusively investigate these possibilities.

An inspection of Figure 1 reveals some other interesting findings. Notable is the high proportion of correct scores for guessing. Since number of answer choices varied between subtests and levels, with four choices the most common format, probability of correct responding by chance alone was estimated at .28. In fact, when students reported guessing, they scored 37% correct. That "guessing" responses scored virtually the same as "seemed right" responses suggests that even when students believe they are guessing, they still have some idea of what the correct answer might be and can use this strategy to advantage. "Seemed right" responses were common on the vocabulary subtests in which students often reported that a particular definition sounded correct, but were otherwise uncertain. Another interesting finding is the high proportion of correct responses when the student reported using outside information or experience. Although content area tests such as science and social studies directly test outside knowledge, reading tests ostensibly are intended to test nothing other than knowledge of the content provided in the passage. So, although use of outside information should not help, in fact, students benefited from the use of such information. (It should be noted, however, that when students referred to the passage, they scored even higher.) What is surprising is that students were able to use outside information as effectively as they did. This finding underlines the problems in "passage independence" of reading comprehension items so well investigated by researchers such as Tuinman (1973-1974).

Level of Confidence as a Variable

Students had a reasonably good idea of whether they had answered a test question correctly or not. When students reported being "very sure" their answer was correct, they were in fact correct 81% of the time. When they reported being "somewhat sure," they were correct only 13% of the time, and when they reported being "not sure", they obtained correct answers in only 7% of the cases. These figures are somewhat misleading, however. If looked at another way, the results seem different: when students answered incorrectly, they also reported being "very sure" the answer was correct in 56% of the cases. Clearly, level of confidence in itself, although related to performance, is not a sufficient check on correctness of a student's work. The relation between confidence to correctness of response was seen to vary widely from student to student, with a median point biserial correlation of .29 ($p > .05$). In many cases, then, other means are necessary for students to assess the correctness of their responses. These means will be described below.

The Cost of Carelessness

In addition to reported test-taking strategies, information was also collected on the degree to which the students attended to distractors and referred to the reading passage on the reading comprehension subtest in choosing their answer. Interestingly, students referred to the reading passage only very rarely, even though when they did refer, they stood a very good chance of answering the question correctly. It was found that when students answered a reading comprehension question incorrectly, in 89% of the cases students had not referred back to the passage which clearly contained the correct answer. This, of course, does not mean that all of these questions could have been answered correctly, but it does appear that reading scores could be much improved by students' increased attention to the passage.

Similarly, a great deal of carelessness was observed in attention to all distractors. When students answered incorrectly, in 40% of the 302 cases they had not read all distractors. Again, this finding does not mean all these questions could have been answered correctly by greater attention to distractors, but the score could almost certainly have been improved by so doing. When students answered questions correctly, they had attended to all distractors in 73% of the 577 cases. It does appear, then, that test performance can be improved through greater attention to distractors.

Another surprising finding was the relatively small effect of reading errors. Although performance was clearly impaired when students misread a word of key importance (see Figure 1), misreading words in general had less detrimental effect than might be expected. When one or more words in stem or distractor were misread, the proportion of items answered correctly (58% of 293) was still quite high. Clearly, many students have developed strategies for coping with words they cannot read. It seems important, then that students be reminded not to "give up" if they cannot read every word. As seen in the present investigation, students are often able to answer correctly even though they were not able to read every word.

One final finding concerning carelessness can be reported. All examiners noted the extent to which students had attended to the wrong stimulus in the "word study skills" subtest. In this subtest, students are given a word with an underlined sound, and asked to find the same sound in one of three distractors. For example, in the following problem:

Prize

(a) prince

(b) size

(c) seven

the correct answer is (b) since the "z" in "size" has the same sound as the underlined "z" in prize. What was surprising to the present investigators is the fact that students so often attended to the wrong stimulus, for example, the initial "pr" in the above question. Although exact incidence of these errors cannot be given, their consistent occurrence seems to imply that teachers should stress the importance of attending to the underlined sound only.

Conclusions

The results of this study have demonstrated that students do employ specific strategies to cope with test item ambiguity and with indecision or lack of knowledge in selecting correct answers. Important implications can be drawn from these findings which have a direct bearing on student performance during testing. To attain the most correct answers, students should employ the strategies listed below:

1. Never skip an answer.
2. Be certain to attend to all distractors and refer to the reading passage, even if you are "very sure" your answer is correct.
3. If you are having great difficulty reading a passage, read the questions and try to answer them anyway. Often, your own knowledge can help you choose an answer. If you have difficulty with some words in the question or distractors, answer anyway and base your answers on the words you can read.
4. If you have attended to all parts of a passage and test question and still do not know an answer, there is still a good chance of getting the correct answer if you guess.
5. Be certain you are attending to the appropriate stimulus, such as the underlined sound in a "word study skills" subtest. As in other subtests, wrong answer choices are given which may look correct at first glance.
6. Make sure you answer every item, even if you must hurry and guess a lot near the end. You will probably get some of the answers correct.

Given the results of past research (Bangert, Kulik, & Kulik, 1983), it is likely that to significantly affect test performance, a teacher will have to do more than simply read the above points to students. Examples and practice activites will help develop these "test-taking" skills.

These findings are of interest to special education, particularly the area of learning disabilities. Many children are referred for special class placement on the basis of deficiencies seen in standardized reading tests. Special education is often quite beneficial to students who clearly need it, but before taking such a dramatic step, it should be known for certain that the student's score reflects the best abilities of the student, rather than a problem with test-taking in general.

Overall, the present investigation indicated that a range of abilities exists in test-taking skills, as it does in other areas. The specific skills observed in efficient students taking a reading test should be practiced by all students, if tests are to be as valid as possible. If test taking skills are incorporated in general test administration procedures, it appears that maximum benefit can be derived from the use of standardized reading tests.

References

Bangert, R. L., Kulik, J. A., & Kulik, C. C. Effects of coaching programs on achievement test scores. Review of Educational Research, 1983, 53, 571-585.

Haney, W., & Scott, L. Talking with children about tests: A pilot study of test item ambiguity. National Consortium on Testing Staff Circular No. 7. Cambridge, Mass.: The Huron Institute, 1980.

Howell, K. W. Evaluating Exceptional Children. Columbus, Ohio: Merrill, 1979.

Millman, J., Bishop, C. H., & Ebel, R. An analysis of test wiseness. Educational and Psychological Measurement, 1965, 25, 707-726.

Spache, G. Identifying and diagnosing reading difficulties. Boston: Allyn & Bacon, 1976.

Slakter, M. J., Koehler, R. A., & Hampton, S. H. Grade level, sex, and selected aspects of test-wiseness. Journal of Educational Measurement, 1970, 7, 119-122.

Taylor, C., & White, K. R. The effects of reinforcement and training on group standardized test behavior. Journal of Educational Measurement, 1982, 19, 199-210.

Tuinman, J. J. Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 1973-1974, 2, 206-223.

Footnotes

[1] The authors would like to thank Dr. Ginger Rhode and Judy Johnson, as well as Dr. Jay Monson, acting director, and the staff of the Edith Bowen school, particularly Dorothy Dobson and Lou Anderson, for their valuable assistance with this project. The authors would also like to thank Ursula Pimentel and Marilyn Tinnakul for typing the manuscript. Address requests for reprints to Thomas E. Scruggs, Ph.D., Exceptional Child Center, UMC 68, Utah State University, Logan, Utah, 84322.

[2] A point-biserial, rather than a Spearman correlation of ranks coefficient, was computed out of concern for the necessarily high number of ties resulting in computing a rank correlation with binary data. The obtained Spearman coefficient, .55, however, differed by only one point from the obtained point biserial of .54.

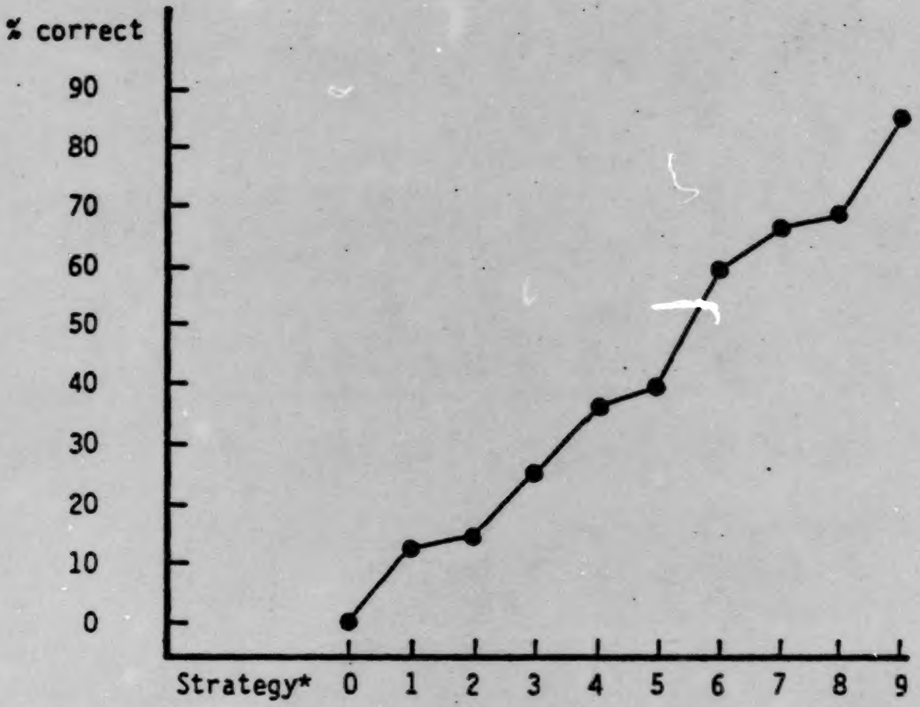
Table 1

Frequencies and Percent of Strategies Employed

Strategy level	Frequency	Percent
0. Skipped Item	9	1.0
1. Misread Keyword	23	2.6
2. Faulty Reasoning	38	4.3
3. Did Not Follow Directions	7	0.8
4. "Seemed Right"	92	10.5
5. General	127	14.4
6. Used External Evidence	21	2.4
7. Eliminated	45	5.1
8. Referred To Passage	59	6.7
9. Clearly "Knew"	458	52.1

Figure Caption

Figure 1. Proportion correct by strategy used.



* Strategy Classifications:

0. Skipped Item
1. Misread Keyword
2. Faulty Reasoning
3. Didn't Follow Directions
4. "Seemed Right"
5. Guessed
6. Used External Evidence
7. Eliminated
8. Referred to Passage
9. Clearly "Knew"

Developmental Aspects of Test-Wiseness for Absurd Options: Elementary School Children

Thomas E. Scruggs
Exceptional Child Center
Utah State University

Running head: Developmental Aspects

Abstract

Twenty-eight students from grades 1 through 5 were administered a test of test-wiseness for absurd options. Results suggested that a developmental trend may exist in test-wiseness for elementary-age school children.

Developmental Aspects of Test-Wiseness for Absurd Options: Elementary School Children

First discussed by Thorndike in 1951, test-wiseness (TW) was described in detail by Millman, Bishop, and Ebel (1965), and defined as "a subject's capacity to utilize the characteristics and formats of the test and/or test-taking situation to receive a high score" (p. 707). They further described TW as "logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures" (Millman et al., 1965, p. 707). Ebel (1965) has suggested that error in measurement is more likely to be obtained from students low in test-taking skills. The student low in TW, therefore, may be more of a measurement problem than the student high in TW (Slakter, Koehler, & Hampton, 1970b).

Some investigations have indicated that TW has a developmental component; that is, that TW increases with age. Slakter, Koehler, and Hampton (1970a) administered a measure of TW to students from grades 5-11 and found a significant overall linear trend for grade level. Crehan, Koehler, and Slakter (1974) administered a TW test to students in grades 7 through 11, and a follow-up test to the same students two years later. Increases over all intervals except grades 9 to 11 were found. In a second follow-up of the same students, Crehan, Gross, Koehler, and Slakter (1978) replicated the previous findings and concluded that although TW increases by grade, large individual differences exist within grade levels.

Although the above investigations provide strong support for a developmental component of TW in the secondary grades, as yet no investigation has evaluated the developmental nature of TW in the elementary grades. The present investigation is intended to address this question.

Method

Subjects were 28 elementary school-age children attending summer classes prior to entering grades 1 through 5 in a western rural community. Students (1 first grader, 9 second graders, 11 third graders, 2 fourth graders, and 4 fifth graders) were selected from both remedial and "enrichment" classes so that a variety of ability levels was sampled.

Students were seen individually by one of four examiners. First, they were administered a five-item test of TW. This test was developed to measure the ability of students to eliminate options known to be incorrect (corresponding to the Millman et al., 1965 TW category I-D-1, absurd options). For example, one of the items was the following:

> Good airplane pilots must be able to ________ quickly in an emergency.
>
> 1. fall asleep 3. sturnate
>
> 2. scream 4. thing

Students were orally provided with words they were unable to read. Since it was thought that evidence of TW would be more subtle in an elementary school population than it was in studies of secondary students, some departures were made from the procedures of Crehan et al. (1974). First, students were directly questioned regarding the reasons for their answer choices following completion of the test. Second, students were scored as reporting no elimination strategies (0), or reporting one or more strategies (1), regardless of the "correctness" of their answer to each test question.

Results and Discussion

A point-biserial correlation was computed between entering grade level of student and presence or absence of reported elimination strategies. The resulting coefficient, .44, was statistically significant ($p < .02$) and represented a moderate relation between grade level of student and reported use of elimination strategies, accounting for approximately 20% of total variance. Proportion of students reporting use of elimination strategies by grade level is given in Figure 1.

Insert Figure 1 about here

Thus, it appears that a developmental trend in one aspect of TW can be observed in children of elementary school age, and that this trend is similar to that seen in older students. These findings must be interpreted with caution, however, due to the limited sample size, as well as the fact that only one aspect of TW was measured. Although further research is needed, the results of this preliminary investigation suggest that students begin to learn TW skills as early as the primary grades, and that these skills continue to improve with age.

References

Crehan, K. D., Gross, L. J., Koehler, R. A., & Slakter, M. J. Developmental aspects of test-wiseness. Educational Research Quarterly, 1978, 3, 40-44.

Crehan, K. D., Koehler, R. A., & Slakter, M. J. Longitudinal studies of test-wiseness. Journal of Educational Measurement, 1974, 11, 209-212.

Ebel, R. L. Measuring educational achievement. New Jersey: Prentice-Hall, 1965.

Millman, J., Bishop, H., & Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 1965, 25, 707-726.

Slakter, M. J., Koehler, R. A., & Hampton, S. H. Grade level, sex, and selected aspects of test-wiseness. Journal of Educational Measurement, 1970, 7, 119-122. (a)

Slakter, M. J., Koehler, R. A., & Hampton, S. H. Learning test-wiseness by programmed texts. Journal of Educational Measurement, 1970, 7, 247-254. (b)

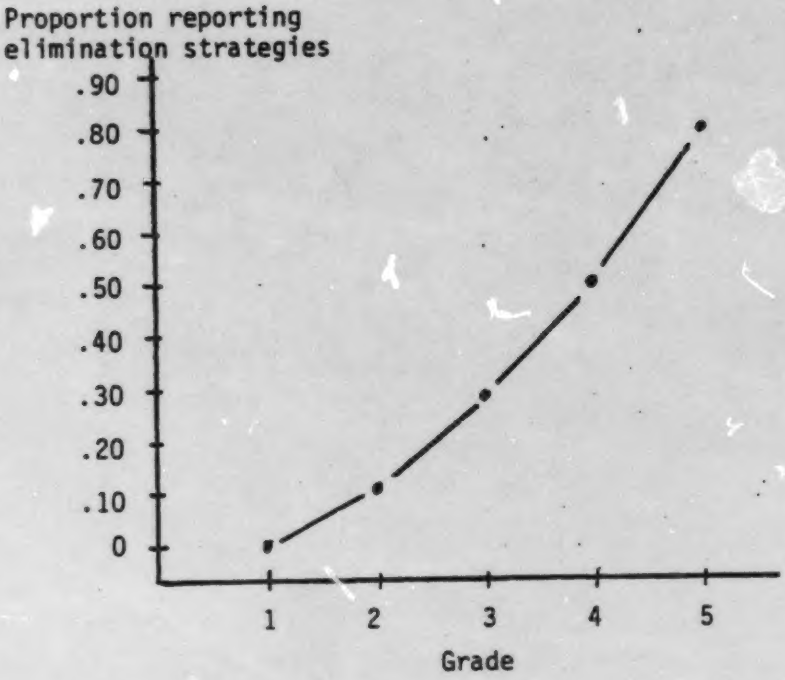
Thorndike, R. L. Reliability. In E. F. Linguist (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1951.

Footnote

[1] The author would like to thank Karla Bennion, Steve Lifson, Dr. Jay Monson and the staff of the Edith Bowen School for their assistance on this project.

Figure Caption

Figure 1. Proportion of students reporting elimination strategies by grade level.


Proportion reporting
elimination strategies
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
1
2
3
4
5
Grade

Format Changes in Reading Achievement Tests:
Implications for Teachers

Karla Bennion

Steve Lifson

Thomas E. Scruggs

Exceptional Child Center

Utah State University

Running head: FORMAT CHANGES

Summary

It has been seen that children's scores on reading achievement tests vary not only with knowledge of content but also with the differing formats of test items. Teachers working with learning disabled children or children with attention problems may wish to choose standardized tests with fewer rather than more format changes. The present study evaluated the number of format and direction changes, across tests and grade levels, of the major elementary standardized reading achievement tests. The number of format changes varies from one change every 3.2 minutes on the California Achievement Test Level 13 to one change every 40 minutes on the upper levels of the Metropolitan Achievement Test. Teachers may wish to take this into account when considering standardized reading achievement tests for their students.

Format Changes in Reading Achievement Tests

It has been seen that the format of achievement test items has an effect on children's test scores (Benson & Crocker, 1979; Carcelli & White, 1981). In one study of reading achievement, children's responses to items with the same content but in different format varied from 45% to 92% correct (White, Carcelli, & Taylor, 1981). Children in grades lower than the fifth grade have attained significantly lower test scores when the major format change of using a separate answer sheet is introduced (Harcourt Brace Jovanovich, 1973; Ramseyer & Cashen, 1971; Cashen & Ramseyer, 1969). Learning disabled children, children with attention problems, and children functioning below grade level may be even more adversely affected by format changes.

Given the extent to which different formats inhibit correct responding, and the lesser ability of children at earlier developmental stages to adjust to major format changes, teachers of such students may wish to consider a reading achievement test with less frequent rather than more frequent format changes. Teachers will prefer to use tests on which a student's scores are affected only by knowledge of content, not the ability to adjust quickly to format changes. Since format has been shown to be a variable influencing test performance, this investigation intended to compare the number of format changes, across tests and grade levels, of the major elementary standardized reading achievement tests.

Procedure

Reading subtests of the following standardized tests were analyzed for format changes: the Stanford Achievement Test (SAT) levels Primary 1, Primary 2, Primary 3, Intermediate 1, Intermediate 2; the California Achievement Tests (CAT) levels 10-19; the Metropolitan Achievement Tests (MAT) levels Primary 1, Primary 2, Primary 3, and Elementary; the Iowa Tests of Basic Skills (ITBS) levels 7-14; the Comprehensive Tests of Basic Skills (CTBS) levels A-G; and the SRA Achievement Series levels A-D.

A format change was defined as a variation in the number of options per item, a change from column to row or row to column, a change in either stem or options from word to picture to passage to question to cloze item. Comparisons across tests and grade levels were made by dividing the time allowed by the number of formats in the test. For example, 20 minutes/4 formats means that in this case, there is a format change every 5 minutes. Interrater agreement was calculated at 91%.

Results

The standardized test with the least number of format or direction changes is the Metropolitan Achievement Test, which averages one format change every 27 minutes. The MAT upper levels have only one change every 40 minutes. The test with the greatest number of changes is the California Achievement Test, with a format change every 9.1 minutes. The CAT level 13 for second and third graders has a format change every 3.2 minutes. The results for all tests and all levels are presented in Table 1.

Insert Table 1 about here

The mean of the format changes across grade levels varies from one change every 8.8 minutes at grades 2-3 to one change every 15.7 minutes at grades 5-7. These results are summarized graphically in Figure 1.

Insert Figure 1 about here

Discussion

Children's test scores vary not only with knowledge of content, but also with the differing formats of test items. Teachers of children with difficulties may wish to consider standardized tests with fewer rather than more format changes. The number of format changes on the major standardized reading achievement tests varies from one change every 3.2 minutes on the CAT level 13 to one change every 40 minutes on the upper levels of the MAT. If a teacher suspects that students have difficulty adjusting to new formats, she or he will prefer to use a test which allows a reasonable amount of time before switching to a different format.

Reference Notes

1. Carcelli, L., & White, K. R. Effect of item format on students' math achievement test scores. Unpublished manuscript, Utah State University, 1981.
2. White, K. R., Carcelli, L., & Taylor, C. Effect of item format on students' reading achievement test scores. Unpublished manuscript, Utah State University, 1981.

References

Benson, B., & Crocker, L. The effects of item format and reading ability on objective test performance: A question of validity. Educational and Psychological Measurement, 1979, 39, 381-387.

Cashen, V. M., & Ramseyer, G. C. The use of separate answer sheets by primary age children. Journal of Educational Measurement, 1969, 6, 155-158.

Harcourt Brace Jovanovich, Test Department. The effect of separate answer document use on achievement test performance of grade 3 and 4 pupils. Special Report No. 24, June 1973.

Ramseyer, G. C., & Cashen, V. M. The effects of practice sessions on the use of separate answer sheets by first and second graders. Journal of Educational Measurement, 1971, 8, 177-181.

Test References

California Achievement Tests, Levels 10-19, Form C. CTB/McGraw-Hill, Monterey, California, 1977.

Comprehensive Tests of Basic Skills, Levels A-G, Form U. CTB/McGraw-Hill, Monterey, California, 1981.

Iowa Tests of Basic Skills, Levels 7-14, Form 7. A. N. Hieronymus, E. F. Lindquist, H. D. Hoover, et al. Houghton Mifflin Co., The University of Iowa, 1980.

Metropolitan Achievement Tests, Levels P1-E1, Form JS. G. A. Prescott, I. H. Balow, T. P. Hogan, & R. C. Farr. Harcourt Brace Jovanovich, New York, 1978.

Stanford Achievement Test, Levels P1-I2, Form # . E. F. Gardner, H. C. Rudman, B. Karlsen, & J. C. Merwin. Harcourt Brace Jovanovich, New York, 1982.

SRA Achievement Series, Levels A-D, Form 1. Science Research Associates, Chicago, 1978.

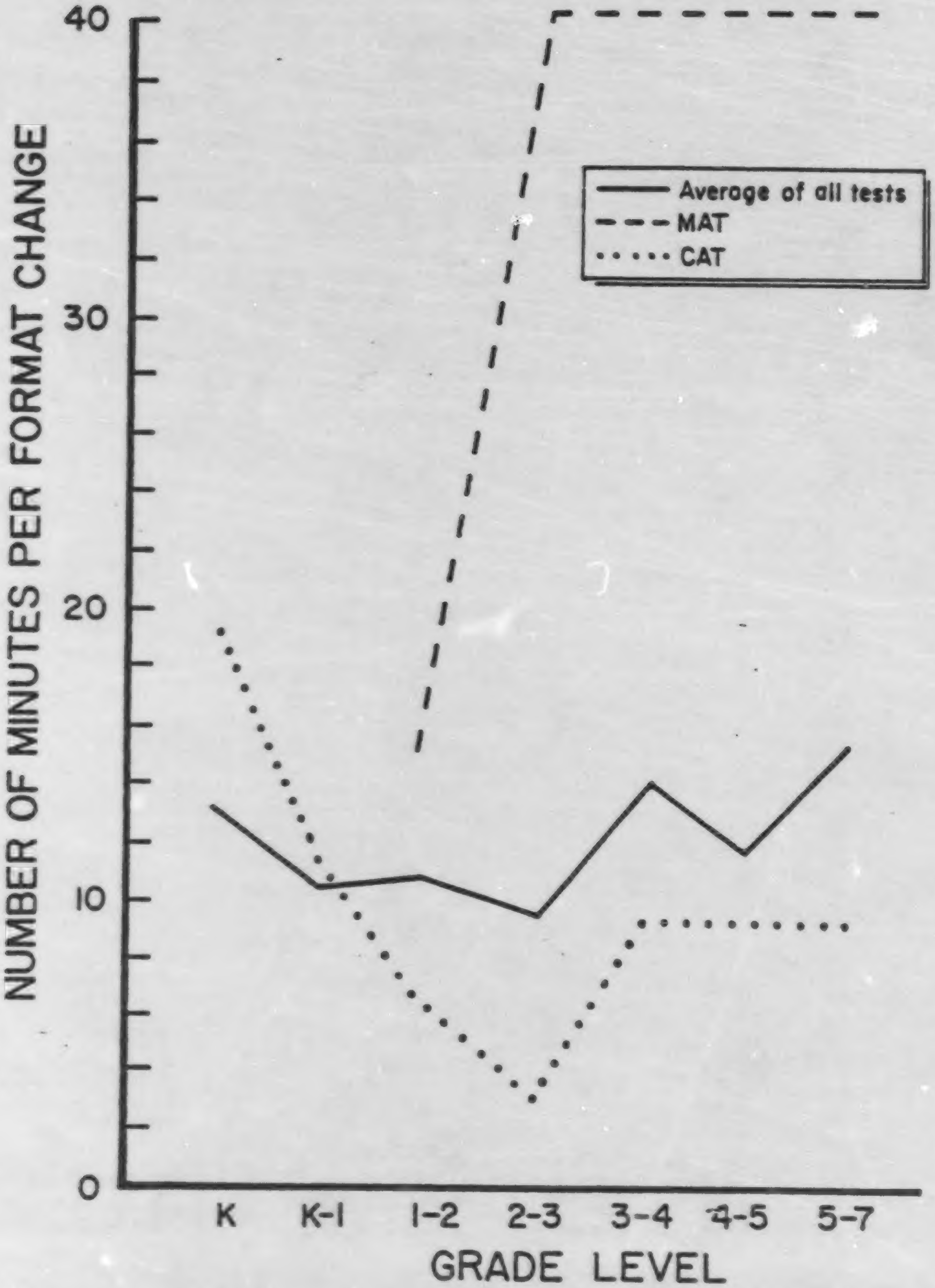
Table 1

Minutes Per Format and Direction Changes

Test	Level	# Minutes/format change
CAT	10	19.3
	11	11.4
	12	6.1
	13	3.2
	14-19	9.0
CTBS	A	7.6
	B	7.5
	C	8.1
	D	8.0
	E	8.8
	F	6.7
	G	6.7
ITBS	7	6.8
	8	6.2
	9-14	19.0
MAT	P1	15.0
	P2	40.0
	E1	40.0
	Int	40.0
SAT	P1	21.2
	P2	15.0
	P3	20.0
	I1	21.3
	I2	21.3
SRA	A	13.9
	B	16.4
	C	14.2
	D	24.0

Figure Caption

Figure 1. Number of minutes per format change.


NUMBER OF MINUTES PER FORMAT CHANGE
0
10
20
30
40
K
K-1
1-2
2-3
3-4
4-5
5-7
GRADE LEVEL
Average of all tests
MAT
CAT

Are Learning Disabled Students "Test-Wise?": An Inquiry into Reading Comprehension Test Items

Thomas E. Scruggs

Steve Lifson

Exceptional Child Center[1]

Running head: READING COMPREHENSION TESTS

Abstract

Previous research has indicated that students in many cases can answer reading comprehension test questions correctly without having read the accompanying passage. The present research compared, in two experiments, the ability of learning disabled (LD) students and more typical age peers to answer such reading comprehension questions presented independently of reading passages. In Experiment 1, LD students scored appreciably lower under conditions resembling standardized administration procedures. In Experiment 2, reading decoding ability was controlled for; however, the performance differential remained the same. Results suggested a relative deficiency on the part of LD students with respect to reasoning strategies and test-taking skills. In addition, the validity of some tests of "reading comprehension" was discussed.

Are Learning Disabled Students "Test-Wise?": An Inquiry into Reading Comprehension Test Items

For many years, there has been some argument over what reading comprehension tests "really" measure (e.g., Thorndike, 1973-1974). The most commonly observed standardized reading comprehension item format consists of a passage and a number of associated multiple choice questions. Reading and understanding the passage is assumed to be a necessary pre-condition to correctly answering the questions. After examining the literature, however, one is forced to question the assumption of question dependence on the stimulus passage. Preston (1964) found that college students were able to answer reading comprehension items with the passages blacked out at a rate significantly above chance. Tuinman (1973-1974) administered five major tests to 9,451 elementary-level students under several conditions. Students in the no passage condition (relevant passage had been blacked out) on the average achieved only 30% fewer correct answers than subjects in the passage-in condition. Similar results were obtained by Pyrczak (1972, 1974, 1975, 1976) and Bickley, Weaver, and Ford (1968). A follow-up study of passage independence by Lifson, Scruggs, and Bennion (1984) revealed that passage-independent items are still quite common in elementary level achievement tests. College undergraduates were able to answer 75%, or almost 12 of 16 questions on the Stanford

Achievement Test, Level P-3, without reading the associated passages. This is considerably above chance.

Scruggs, Bennion, and Lifson (in press) interviewed elementary age students regarding their responses on a reading comprehension test. They found that students often chose their answers based upon their own prior knowledge, rather than content of the reading passage. When students reported using such prior information, they answered correctly in over 60% of the cases.

Reading comprehension items which are independent of the associated passage can be answered on the basis of the following: (a) general knowledge, (b) interrelatedness of the questions on a particular passage, and (c) faulty item construction, i.e., keyed option is twice as long or more precisely stated (Pyrczak, 1975). In the first two cases, the presence of enough information in the question stem to identify the topic is an important factor (e.g., "Which of the following statements is NOT true of penguins?"). Such a stem may render a question answerable in terms of information already available to the examinee, and provide clues to the answers of related questions about the same passage that lack such information in the stem ("This passage is about: a) birds of South America, *b) birds of the Antarctic . . . etc.). These clues which individuals apply to a testing situation to maximize their scores, correspond to Millman, Bishop, and Ebel's (1965) criteria of test-taking skills, or "test wiseness."

While test constructors may be able to point to high validity coefficients for their reading comprehension tests and subtests, an important question arises concerning whether all students are equally able to answer questions with the above mentioned characteristics without reading the passage. Are some groups of students at a relative advantage/disadvantage in ability to answer these questions without reading the passage? To answer this question a group of students classified as learning disabled (LD) and a group of regular classroom students were administered a selection of multiple choice reading comprehension questions with the relevant passages removed. The conditions of this experiment were meant to resemble those of a normal testing situation-i.e., students were required to read the questions without assistance. This did not permit us to determine the extent to which any observed differences between the regular and LD students were due to reasoning or variations in general knowledge between the two groups or simply reflected a difference in reading ability. To address this issue, a second experiment was performed to see if similar differences could be found when word reading was controlled for.

Experiment 1

Method

Subjects and Materials

Subjects consisted of 67 regular classroom and resource room third grade students selected from several elementary schools in a western rural area. Of these subjects, 52 were regular classroom students and 15 were classified as LD by P.L. 94-142 and local criteria, which included a 40% discrepancy between actual and expected performance in two areas of academic functioning. The average grade equivalent of the total reading score of the non-LD students on the Comprehensive Test of Basic Skills (CTBS) was 3.4 (SD=.8), while the average CTBS total reading score for the LD students was 2.1 (SD=.5).

Fourteen multiple choice reading comprehension questions without the accompanying passages were selected for this task. Items were drawn from the Stanford Achievement Test, Level P-3, Form E (1982). Items had been chosen to represent questions thought by the author to be answerable in terms of: (a) the general knowledge of the test taker, and (b) the degree to which the interrelatedness of the items served as a cue to the answers. These items were taken from the Lifson et al. (1984) study, in which students' ability to answer these questions had been documented. The items were kept in clusters which belonged together in terms of association with a particular passage.

Procedure

Treatment was administered in regular instructional groupings. Materials were passed out and all students were told that they were about to take a reading test for which they would not be shown the accompanying reading passages, but that they should try their best to answer all questions. No time limit was imposed upon the task.

Results

The regular classroom group answered correctly approximately 55% of the questions, for mean score of 7.8 (SD=1.96). This score was significantly above a chance score of 3.5 ($\underline{t}(102) = 11.27$, $p<.001$). In contrast, the LD students answered correctly only 35% of the questions, for a mean score of 4.9, only slightly higher than chance ($\underline{t}(28) = 1.77$, ns). The obtained score of the non-LD group was significantly higher than the LD group ($\underline{t}(65) = 4.91$, $p<.001$).

Discussion

The present findings suggest that regular classroom students are able to recognize and make use of cues in testing situations in order to increase their scores, even when reading passages are deleted, and "reading comprehension" supposedly cannot be measured. Apparently, LD students are not able to benefit equally from these cues. Since neither group should have scored above chance on a reading comprehension test with the reading passages deleted, it is possible that a certain amount of bias exists against children with learning disabilities on some standardized tests of reading comprehension. Students in regular classes when unable to read or otherwise obtain meaning from reading passages are still able to answer correctly comprehension questions. Students with learning disabilities, however, do not seem to have these skills, and are thereby punished twice for a reading handicap: once for being less able to read and comprehend the passage, and a second time for being unable to "second guess" test questions, as their nonhandicapped peers are apparently able to do.

One possible explanation for this discrepancy between LD and regular classroom students is that LD students are simply less able to read (decode) the questions, and for that reason are less able to outguess the test. That is, LD students are less deficient in "test taking skills" than they are in reading ability. In order to address this question, a second experiment was designed, in which ability to read would be controlled for. Although the conditions in this experiment could not parallel those of standardized test procedures, they did allow for an assessment of the extent to which differential scores are attributable to lower reading skills, or to lower levels of "test-wiseness."

Experiment 2

Method

Subjects and Materials

The 42 subjects who participated in this investigation were different students drawn from the same population as those of Experiment 1, and consisted of 27 regular classroom third grade students and 15 third grade children classified as LD by P.L. 94-

142 and local district criteria. Mean grade equivalent for the non-LD group (CTBS total reading) was 3.6 (SD=.9), and 1.9 (SD=.4) for the non-LD group. Materials were 14 items drawn from the Stanford Achievement Test, level P3, Form F, and were chosen on the same basis as those used in Experiment 1. Pages of the test were again left intact with questions left in the original order and the passages themselves blacked out during the copying process.

Procedure

Students were informed by their teacher that they were about to take a reading test without reading the corresponding passages. They were told to listen while the teacher read each item, and then answer the items.

Results and Discussion

The students in regular classrooms answered correctly 65% of the fourteen items, for a mean score of 9.14 (SD=1.8). The LD students, on the other hand, answered correctly only 45% of the items, for a mean score of 6.33 (SD=1.8). Although both obtained scores are well above chance, ($t(52) = 12.02$, and $t(28) = 4.325$, ps<.001, for regular classroom and LD students, respectively), the regular classroom group maintained its advantage over the LD students, $t(40)=4.87$, p<.001. The results suggest that learning disabled students are less likely to apply test-taking strategies to reading comprehension questions to a degree of efficiency similar to their non-LD counterparts.

General Discussion

In Experiment 1, regular third grade classroom students were seen consistently to outscore their LD counterparts on a test of reading comprehension questions with corresponding passages deleted, and administered under conditions resembling standardized testing procedures. In Experiment 2, regular class third graders again outscored LD students, under conditions for which reading ability was controlled. The ability of third grade children in these cases to score 55% and 65% correctly on questions which refer to non-existent passages seems remarkable, and brings into question the issue of what some tests of "reading comprehension" are really measuring. Such passage independent items have been thought to assess test-taking skills and in fact have been used as measures of "test-wiseness" (e.g., Derby, 1978). Whatever such tests measure, it is clear that: (a) it is not "reading comprehension," and (b) children classified as LD are at an apparent disadvantage.

An argument can be made that these comparisons are of trivial importance, since in standardized test administration, passages are not deleted; that all children in fact have equal access to passages which contain answers to reading comprehension questions. Although this argument has a certain face validity, some problems remain. First, since non-LD students can score so high on such items without reading the passages, the extent to which scores are a direct measure of "reading comprehension" seems uncertain. Second, since nearly all such tests are timed, students with incomplete understanding of relevant passages but possessing an ability to "outguess" test questions under time constraints, clearly are at an advantage with respect to students not possessing such an ability. In this case, differences in scores on reading comprehension tests may in fact reflect in part a bias toward students with superior "test-wiseness." As has been seen in the present experiments, LD students may well find themselves on the negative side of any such bias.

The extent to which LD and their non-LD counterparts differ on the present measures appears to have surprisingly little to do with reading ability. Although both groups gained when reading (decoding) ability was controlled for, each group was seen to exhibit the same degree of gain, amounting to 10 percentage points for each group. Reported t values in Experiments 1 and 2 remained virtually identical. It seems clear, then, that much of the observed performance difference in Experiment 1 was due to skills other than reading ability, or "reading comprehension." Possibly, relative deficits in vocabulary knowledge account for some of these differences. What also may be a factor is a differential ability to respond to specific cues in the test-taking situation.

Two steps may be taken to help alleviate this potential source of bias. First, achievement tests should be revised so that reading comprehension tests directly assess comprehension of the provided passage. In fact, an informal review by the present authors of the major achievement tests indicates that many achievement test questions appear to be much less "passage independent" since the work of Tuinman (1973-1974) and others of a decade ago. Second, it seems possible that at least some of these "test-taking skills" can be trained, and that this training may do much to correct this apparent disadvantage. The authors are at present investigating the effectiveness of such training (Taylor & Scruggs, 1983). Although such improved scores on tests may not necessarily reflect increased achievement, these scores could reflect more accurately achievement gains students have made, as evaluated by standardized achievement tests.

References

Bickley, A. C., Weaver, W. W., & Ford, F. G. (1968) Information removed from multiple-choice item responses by selected grammatical categories. Psychological Reports, 23, 613-614.

Derby, T. C. (1978). The effects of instruction in selected aspects of test-wiseness on the administration of standardized test items in the upper elementary schools. Unpublished doctoral dissertation, Southern Illinois University, Carbondale.

Lifson, S. A., Scruggs, T. E., & Bennion, K. E. (1984). Passage independence in reading achievement tests: A follow-up. Perceptual and Motor Skills, 58, 945-946.

Preston, R. C. (1964) Ability of students to identify correct responses before reading. Journal of Educational Research, 58, 181-183.

Pyrczak, F. (1972). Objective evaluation of the quality of multiple-choice test items designed to measure comprehension of reading passages. Reading Research Quarterly, 8, 62-71.

Pyrczak, F. (1974). Passage-dependence of items designed to measure the ability to identify the main ideas of paragraphs: Implications for validity. Educational and Psychological Measurement, 34, 34-348.

Pyrczak, F. (1975). Passage-dependence of reading comprehension questions: Examples. Journal of Reading, 19, 308-311.

Pyrczak, F. (1976). Context-indendence of items designed to measure the ability to derive the meanings of words from their context. Educational and Psychological Measurement, 36, 919-924.

Scruggs, T. E., Bennion, K. E., & Lifson, S. A. (in press). An analysis of children's strategy use on reading achivement tests. Elementary School Journal.

Taylor, C., & Scruggs, T. E. Improving the test-taking skills of learning disabled and behaviorally disordered elementary school children. Exceptional Children, 1983, 50, 277.

Thorndike, R. L. (1973-1974). Reading as reasoning. Reading Research Quarterly, 9, 135-147.

Tuinman, J. J. (1973-1974). Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 9, 206-223.

Footnote

[1] This research was supported in part by a grant from the U.S. Department of Education. The authors would like to thank the excellent teachers of Cache Valley, Utah, for their assistance with this project: Marian Innocenti, Brenda Neiderhauser, Bonnie Olsen, Loila Anderson, and Edna Eams were particularly helpful. The authors would also like to thank Jill Barry, Ursula Pimentel, and Marilyn Tinnakul for their assistance in the preparation of this manuscript. Address requests for reprints to Thomas E. Scruggs, Exceptional Child Center, UMC 68, Utah State University, Logan, Utah 84322.

PASSAGE INDEPENDENCE IN READING ACHIEVEMENT TESTS: A FOLLOW-UP[1]

STEVE LIFSON, THOMAS E. SCRUGGS, AND KARLA BENNION

Utah State University

Summary.—38 college undergraduates were administered reading-comprehension items from a major standardized achievement test with corresponding passages deleted. Analysis indicated that, after 20 years of similar research findings, highly passage-independent items still occur on major tests.

For almost 20 years, it has been documented that reading-comprehension test items can be answered correctly at above-chance rates without actually reading the relevant passage (Preston, 1964). Pyrczak (1976) mentions several types of items which seem particularly independent of the passage. These types include (a) items that can be answered from the examinee's own knowledge and (b) items about a particular passage that are related to each other in such a way that some items provide clues for other items. Reading-comprehension tests which include such items invite critical attention on the grounds that (a) examinees may have an advantage over those not using these strategies (Pyrczak, 1972) and (b), if a subject uses these principles and skips passages, he invalidates the purpose of the test (Tuinman, 1973-1974). Since an extensive review of the literature has shown no justification for the use of passage-independent items, the question arises as to whether these items still occur in commonly used standardized achievement tests. The present investigation was intended to determine whether such items are still in use.

METHOD

Subjects and Materials

Thirty-eight undergraduate elementary education students at a western university completed 16 multiple-choice reading-comprehension questions without the accompanying passages. The items selected were thought to represent questions that could be answered without having read the accompanying passage. These items were chosen to correspond to Millman, Bishop and Ebel's (1965) categories of test-wiseness strategies involving the general knowledge of the test taker and use of subject matter of neighboring items. The specific effects of these cues, however, were not addressed in this study. The 16 items were taken from the Stanford Achievement Test Form E, Level P-3, from a pool of 60 items. The items were kept in clusters illustrating which belonged together in terms of association with a particular passage.

---

[1] The authors thank Dr. Barnard Hays for his kind and generous assistance with this investigation. Requests for reprints should be addressed to Steve Lifson, Exceptional Child Center, UMC 68, Utah State University, Logan, Utah.

Procedure

The materials were distributed to two sections of a class in teching reading. The students were told: "Today I'm going to give you some reading-comprehension test items ***without*** the passages. It is not expected that you will answer all of the questions correctly; just do your best. Guess if you do not know the answer." No time limit was imposed upon the task.

RESULTS AND DISCUSSION

Analysis indicated that the mean score was 75% correct, with an average mean score of 11.9 of the 16 items. A one-sample t test (Hayes, 1973) confirmed that the obtained scores were significantly different from chance responding ($t = 18.9$, $p < .001$).

Although the items were not randomly selected for this measure, they nevertheless represented 25% of the items included in the reading-comprehension section of the test. Clearly, at least some test developers have done little to alter passage-independent items in light of the research findings of almost two decades. While the effects of the readers' previous knowledge cannot be eliminated, the effects could be minimized by the use of fictional material for the passages with accompanying questions about the activities of an imaginary person. In spite of the reported validity of these items (SRA, 1979), the burden of construct validity rests with the authors of the tests. If some students are able to answer "reading-comprehension" test items correctly without reading the passage, one can question what is being measured.

REFERENCES

GARDNER, E. F., RUDMAN, H .C., KARLSEN, B., & MERWIN, J. C. *Stanford Achievement Test, Form E.* New York: Harcourt, Brace, Jovanovich, 1982.

HAYS, W. L. *Statistics for the social sciences.* New York: Holt, Rinehart & Winston, 1973.

MILLMAN, J., BISHOP, C. H., & EBEL, R. An analysis of test-wiseness. *Educational and Psychological Measurement,* 1965, 25, 707-726.

PRESTON, R. C. Ability of students to identify correct responses before reading. *Journal of Educational Research,* 1964, 58, 181-183.

PYRCZAK, F. Context-independence of items designed to measure the ability to derive the meanings of words from their context. *Educational and Psychological Measurement,* 1976, 36, 919-924.

SCIENCE RESEARCH ASSOCIATES, INC. *SRA achievement series answer keys, norms, and conversion tables, level C/forms 1 and 2.* Chicago, IL: Author, 1979.

TUINMAN, J. J. Determining the passage dependency of comprehension questions in five major tests. *Reading Research Quarterly,* 1973-1974, 9, 206-223.

Accepted ....

BEST COPY AVAILABLE

Spontaneously Employed Test-Taking Skills of Learning Disabled Students on Reading Achievement Tests[1]

Thomas E. Scruggs, Karla Bennion, and Steve Lifson

Utah State University

Running head: TEST-TAKING SKILLS

Abstract

The present investigation was intended to provide information on the type of strategies employed by learning disabled (LD) students on standardized, group-administered achievement test items. Of particular interest was level of strategy effectiveness and possible differences in strategy use between LD and non-disabled students. Students attending resource rooms and regular third grade classes were administered items from reading achievement tests and interviewed individually concerning the strategies each had employed in answering the questions and level of confidence in each answer. Results indicated that (a) LD students were less likely to report use of appropriate strategies on inferential questions, (b) LD students were less likely to attend carefully to specific format demands, and (c) levels of confidence reported by LD students were inappropriately high.

Spontaneously Employed Test-Taking Skills of Learning Disabled Students on Reading Achievement Tests

Since the seminal work of Millman, Bishop, and Ebel in 1965, concern has been given to the issue of test-taking skills, or "test-wiseness," as a source of measurement error in group-administered achievement tests (Sarnacki, 1979). Defined as "a subject's capacity to utilize the characteristics and formats of the test and/or the test-taking situation to receive a high score" (Millman et al., 1965, p. 707), test-wiseness is said to include such diverse components as guessing, time-using, and deductive reasoning strategies. Given that the effective use of such strategies may have little to do with knowledge of a particular academic content area, individuals or groups of individuals lacking in these skills may be at a disadvantage. A recently completed meta-analysis, for example, has suggested that under certain circumstances, low-SES students are more likely to benefit from achievement test "coaching" than are higher SES students, which finding implies low-SES students are relatively deficient in the area of test-taking skills (Scruggs, Bennion, & White, 1984).

The present investigation was concerned with the spontaneous use of such strategies by learning disabled (LD) children. Part of a larger investigation involving test-taking skills of exceptional students (Taylor & Scruggs, 1983), the present study had as a goal the identification of possible deficits in test-taking skills on the part of LD children. Such deficits, if uncovered, could be helpful in developing techniques for remediation.

Although much research has been conducted on non-handicapped populations in the area of test-taking skills (see Bangert-Drowns, Kulik, & Kulik, 1983; Sarnacki, 1979; and Scruggs, Bennion, & White, 1984, for reviews), little is known about test-taking skills exhibited by LD children. Scruggs and Lifson (1984) recently investigated the differential ability of LD students to answer "passage independent" reading comprehension test items (i.e., reading comprehension test items for which relevant passages had been omitted). Items were taken from standardized achievement tests known from previous research to be answerable without having read the associated passage (Lifson & Scruggs, in press), and thought to be a good measure of "test-wiseness." In two experiments, non-handicapped children scored 55% and 65% correct on such items, while LD children from the same grade scored much lower, even when word reading ability was controlled. Scruggs and Lifson (1984) argued that such findings also raised the question of what reading comprehension tests "really" measure since no reading comprehension test items should be answerable without having read the associated passage. Scruggs and Lifson concluded that LD children may be at a relative disadvantage with respect to such test-taking skills as guessing, elimination, and deductive reasoning strategies applied to response items.

Scruggs, Lifson, and Bennion (in press) recently employed individual interview techniques to more precisely determine the nature of the strategies spontaneously produced by elementary school children on reading achievement tests. Students representing a wide range of age and ability levels were given reading achievement test items appropriate to each student's reading level. Results indicated that students employed a wide range of strategies on reading achievement tests, far beyond simply "knowing" or "not knowing" the answer, and that the use of these strategies was strongly predictive of performance. These findings provided valuable general information regarding the manner in which children respond to reading achievement test items. However, the diversity of the population in age and achievement level was thought to have obscured observation of specific differences in test-taking skills between age or ability levels. The present investigation, therefore, was intended to determine whether differences in strategy use existed on reading achievement tests between LD and non-disabled students. In this investigation, grade level was held constant and the number of subtests was reduced to two: a "reading comprehension" subtest, in which direct referring, elimination, and deductive reasoning strategies were thought to be important; and a "letter sounds" subtest, in which close attention to format demands was thought to be of importance. In addition, since level of reported confidence was found to be a strong predictor of performance (Scruggs, Bennion, & Lifson, in press), and a prerequisite to strategy monitoring, confidence reports were examined for possible differences between ability groups.

Method

Subjects

Subjects were 32 third grade students attending public schools in a western university community. Twelve were classified as learning disabled (LD) according to local school district criteria, which included a 40% discrepancy between ability and performance in two academic areas, and Public Law 94-142. Twenty were regular classroom students, none of whom were referred for special services and who were thought by their teachers to be functioning within a normal range of performance. Mean grade equivalent for reading comprehension on the Comprehensive Test of Basic Skills (CTBS) was 2.29 (SD=.29), equivalent to a percentile score of approximately 21. Mean grade equivalent on the CTBS for the non-LD students was 3.91 (SD=.89), equivalent to a percentile score of 61. The 16 boys and 16 girls were all 8-9 years old and Caucasian. Sex was evenly represented in LD and non-LD groups.

Materials

Two reading tests were constructed from items taken from the

Stanford Achievement Test. Items were drawn from the Primary 2 battery for the instrument used with the LD group, and the Intermediate 1 level served as the source for the regular classroom group. Each test contained three reading passages with 14 dependent questions (10 content, 4 inference) on each form. Comprehension questions were left in their original order in relation to the passage. The questions were renumbered to avoid gaps where passages did not follow each other sequentially in the original test. In addition, three items from the letter-sound test (level P3) were selected. These consisted of a stimulus word with a letter or letters underlined representing a sound that the student was to identify in the three options given below the stimulus word. These items were selected to include a distractor that closely matched the initial consonants of the stimulus word. For example, in the item:

blind

0 blink
0 nibble
0 leaned

"leaned" is the correct answer, since it contains the same sound as the underlined nd in the stem, and "blink" is the inappropriate distractor, since it contains the same initial consonant blend.

Procedure

Subjects were seen individually by one of two examiners.

They were asked to read the passages and questions aloud and mark answers they thought were correct. They were then told that they would be asked if they were sure/not sure that the answer they had given was correct, and the manner in which they had chosen the particular answer. The subject's response to the questions, "How did you choose that answer?" and "Are you sure or not sure of your answer?" were recorded verbatim on the protocol. Words the experimenters had previously deemed essential to answering the questions (key words) were marked in the examiner's copy of the instrument, and errors in these words were noted as the child read aloud.

Scoring

Test items were scored for correctness, confidence in answer (sure/not sure), and type of strategy reported. Two students from the non-LD group, who had misread more than 25% of the keywords, were excluded from further analysis. The responses given by the subjects were divided into seven logical categories:

1 = Don't know

2 = Guessed

3 = External source of knowledge (e.g., "I know all fish have scales")

4 = Refers to passage (e.g., "I read it")

5 = Quotes directly (e.g., "It says here that . . .")

6 = Eliminates options known to be incorrect

7 = Other reasoning (e.g., "It said comforted in the story. That sort of means relieved.")

Each response was then evaluated in terms of those seven categories. Percent of agreement for scoring was assessed at 100% after each examiner scored 25% of the other examiner's protocols.

Results

A t test applied to percent of keywords read incorrectly indicated that the groups did not differ significantly with respect to reading difficulty, $t(29) = .37$, $p > .20$. Overall, LD students misread 6.6% of (30) total keywords and non-LD students misread 6.75% of (29) keywords.

Proportion correct by collapsed strategy group (inappropriate = strategies 1-3; referring = strategies 4-5; reasoning = strategies 6-7) was computed for item type and student group and is given in Figures 1 and 2.

Insert Figures 1 and 2 about here

Reported strategy data were scored for appropriateness of reported strategy. Strategies were considered appropriate if students reported referring to the passage on a recall question (strategy 4 or 5), or if they reported a reasoning strategy in response to an inferential question (strategy 6 or 7). Proportion of appropriate responses were then entered into a 2 group (LD vs.

non-LD) by 2 item type (direct recall or inferential) analysis of variance (ANOVA) with repeated measures on the item type variable. Because of the unequal group frequencies, a least-squares method of analysis (Winer, 1971) was employed. Significant differences were found for item type, $\underline{F}(1,29) = 9.19$, $\underline{p} < .01$, and for interaction, $\underline{F}(1,27) = 7.58$, $\underline{p} < .05$. Figure 3 depicts graphically the interaction effect. Although both LD and non-LD students reported a high proportion of "referring to text" strategies on recall questions (89% vs. 77%, respectively), large differences emerged in proportion of reasoning strategies applied to inferential questions (39% vs. 70%, respectively). Nonsignificant differences were observed for overall group means, $\underline{F}(1,29) = 1.54$, ns.

Insert Figure 3 about here

Analysis of confidence reports indicated that both groups were similar with respect to reported level of confidence on "referring to passage" strategies with LD students reporting confidence in 85% of the cases and non-LD students reporting confidence in 92% of the cases. These reports were similar to actual performance, with correct scores of 81% and 86% on these items for LD and non-LD groups, respectively. On reasoning strategies, however, a much different picture emerged. Average students were correct on 83% of inferential items, but reported confidence on an average of 71% of the items. The LD students, on the other hand, reported being confident on an average of 95% of the cases, but were in fact correct in only 63% of these cases.

Items on the letter-sound subtest were scored for responses which suggested attention to an inappropriate distractor. This inappropriate distractor took the form of an initial consonant blend present in the stem, but not underlined. Number of inappropriate distractors chosen was compared by group, and differences found to be significant, $\underline{t}(28) = 2.47$, $\underline{p} < .05$. The LD children chose the inappropriate distractor in 52% of the cases, while the non-LD children chose the inappropriate distractor in only 24% of the cases.

Discussion

It has been seen that the present sample of LD third graders, with reading ability controlled for, differed from their more average counterparts with respect to (a) proportion of appropriate reasoning strategies reported for inferential comprehension questions, (b) performance and confidence level for items in which reasoning strategies had been reported, and (c) choice of an inappropriate distractor on a letter-sounds test. On the other hand, LD students did not differ from their more average counterparts with respect to appropriate strategy use on recall items. Generally, this sample of LD children was seen to report fewer reasoning strategies, when appropriate, on reading comprehension test items than did their more average counterparts, and to be less successful on those items for which they did report reasoning strategies. These findings support those reported by Scruggs and Lifson (1984). In that study, LD children were seen to exhibit relatively inferior performance on a test of selected reading comprehension test items for which the relevant passages had been removed, and for which reasoning strategies were thought to be necessary in order to answer the items correctly. The present finding of inappropriately high levels of confidence exhibited by the LD students on items for which reasoning strategies had been applied is supportive of a theory of a developmental deficit in "meta-cognitive abilities" (cf. Torgesen, 1977), in that inappropriately high levels of confidence in task performance are often seen in younger children. This relative deficit on the part of LD children is thought to be a critical one, for ability to evaluate accurately a chosen response is a necessary prerequisite for effective test-taking performance.

That LD students more often attended to an inappropriate distractor may be a function of an attentional deficit (Krupski, 1980) on test format as much as a deficit in phonetic skills. These "test-taking skills" may or may not be subject to simple remediation (Taylor & Scruggs, 1983), but they may reflect a source of measurement error (Millman, Bishop, & Ebel, 1965).

Reading comprehension, clearly, is a construct which seems to resist precise analysis and for which many theoretical orientations exist (Spiro, Bruce, & Brewer, 1980). If one does look at recall and inference as two component parts of reading comprehension, however, it appears from the present investigation that relative strategy and performance deficits exist on the part of LD children on inference questions, but not on recall questions, with reading ability controlled for. To this extent, one could argue that the specific deficits exhibited here reflect problems in reading comprehension itself rather than "test-taking skills," and it does seem likely that strategy training in such areas could reflect improved reading comprehension skills as well as improved test-taking skills, particularly in that selecting and implementing appropriate strategies has been used in research to improve general cognitive functioning (cf. Torgesen & Kail, 1980). In the word study skills subtest, however, the LD students apparently became confused by specific format demands which likely had little to do with the content being tested. Training for this type of strategy deficit, then, would not be expected to bring about a concomitant increase in phonetic analysis skills.

Replication is necessary to further support and refine these findings. The present results suggest that LD children may benefit from specific training in (a) attending to specific format demands, (b) identifying inference questions, and (c) selecting and applying appropriate strategies relevant to those questions.

References

Bangert, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. _Review of Educational Research_, _53_, 571-585.

Krupski, A. (1980). Attention processes: Research, theory, and implications for special education. In B. Keogh (Ed.), _Advances in special education_ (Vol. 1). Greenwich, Connecticut: Jai Press.

Millman, J., Bishop, C. H., & Ebel, R. (1965). _An analysis of test wiseness_. Educational and Psychological Measurement, _25_, 707-726.

Sarnacki, R. E. (1976). An examination of test-wiseness in the cognitive test domain. _Review of Educational Research_, _49_, 252-279.

Scruggs, T. E., Bennion, K., & Lifson, S. (in press). An analysis of children's strategy use on reading achievement tests. _Elementary School Journal_.

Scruggs, T. E., Bennion, K., & White, K. _The effects of coaching on achievement test scores in the elementary grades: A meta-analysis_. Unpublished manuscript, Utah State University.

Scruggs, T. E., & Lifson, S. (1984). _Are learning disabled students 'test-wise?': An inquiry into reading comprehension test items_. Unpublished manuscript, Utah State University.

Scruggs, T. E., & Lifson, S. (in press). Passage independence in reading comprehension items: A follow-up. Perceptual and Motor Skills.

Spiro, R. J., Bruce, B. C., & Brewer, W. F. (1980). Theoretical issues in reading comprehension. Hillsdale, NJ: Erlbaum.

Taylor, C., & Scruggs, T. E. (1983). Research in Progress: Improving the test-taking skills of learning disabled and behaviorally disordered elementary school children. Exceptional Children, 50, 277.

Torgesen, J. K. (1977). Memorization processes in reading disabled children. Journal of Educational Psychology, 69, 571, 578.

Torgesen, J. K., & Kail, R. V. (1980). Memory processes in exceptional children. In B. Keogh (Ed.), Advances in special education (Vol. 1). Greenwich, Connecticut: Jai Press.

Winer, B. J. (1971). Statistical principles and experimental design (2nd Ed.). New York: McGraw-Hill.

Footnote

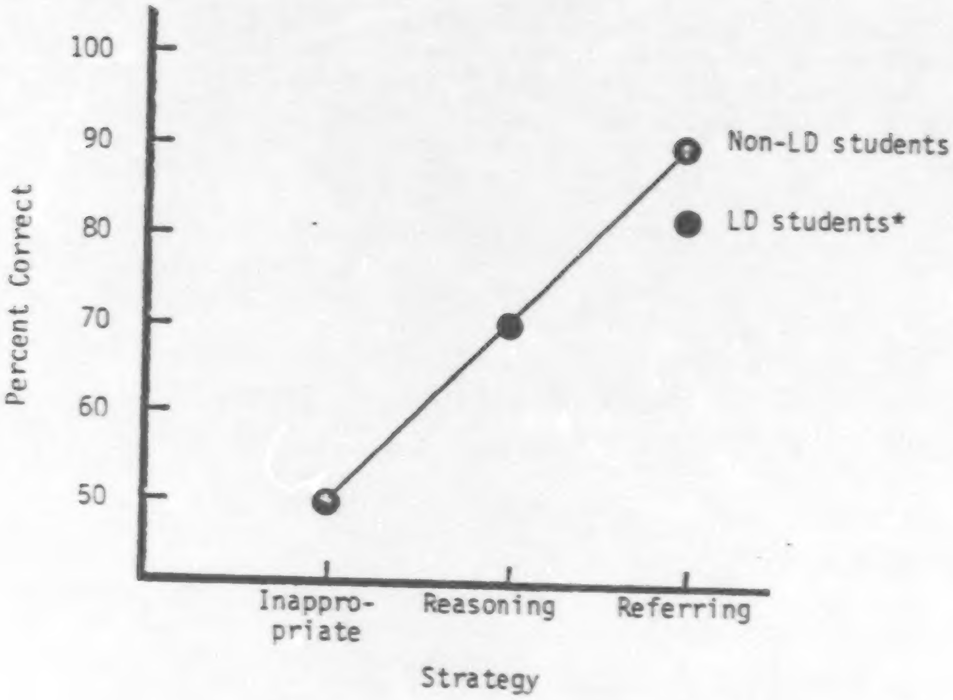
[1] The authors would like to thank Mrs. Bonnie Olsen, as well as Dr. Ted Williams, director, and the staff of the Edith Bowen School, particularly Dorothy Dobson, for their valuable assistance with this project. The authors would also like to thank Marilyn Tinnakul and Jill Barry for typing the manuscript. Address requests for reprints to Thomas E. Scruggs, Ph.D., UMC 68, Utah State University, Logan, Utah 84322.

Figure Captions

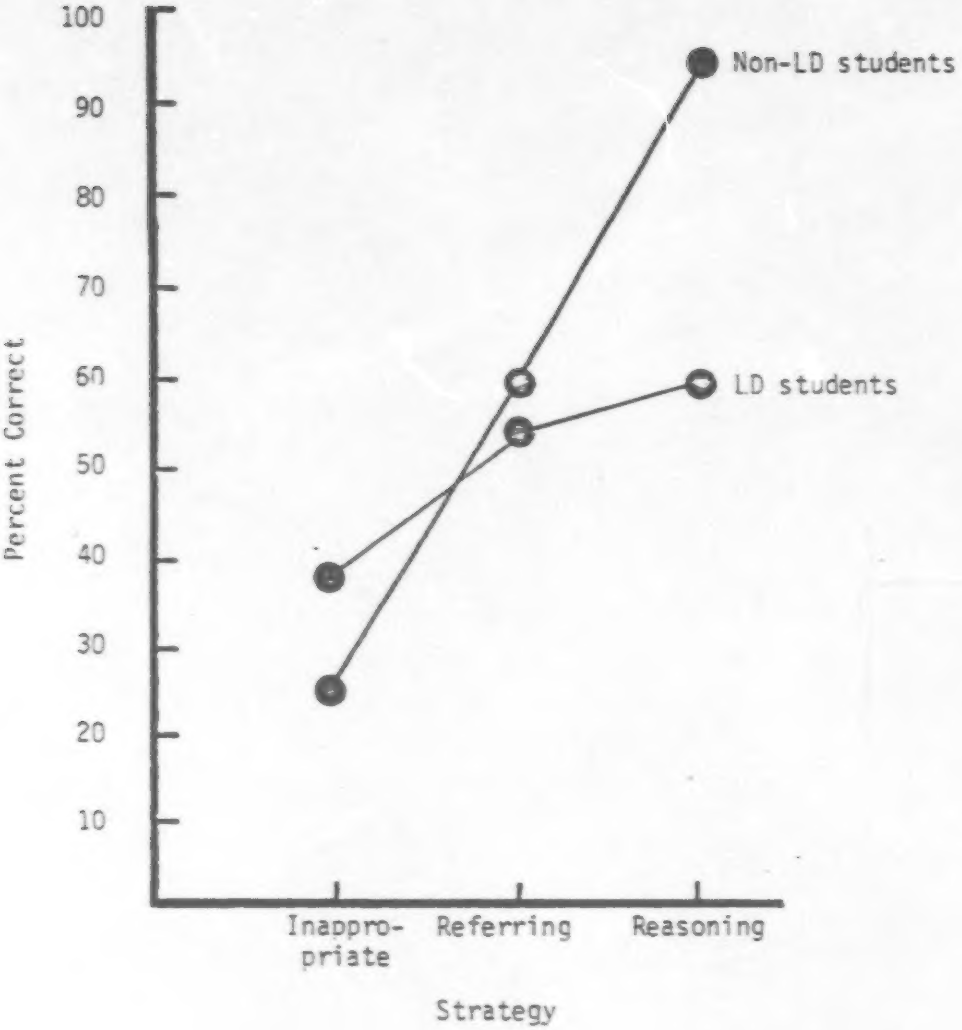
Figure 1. Proportion correct by strategy used on recall item.

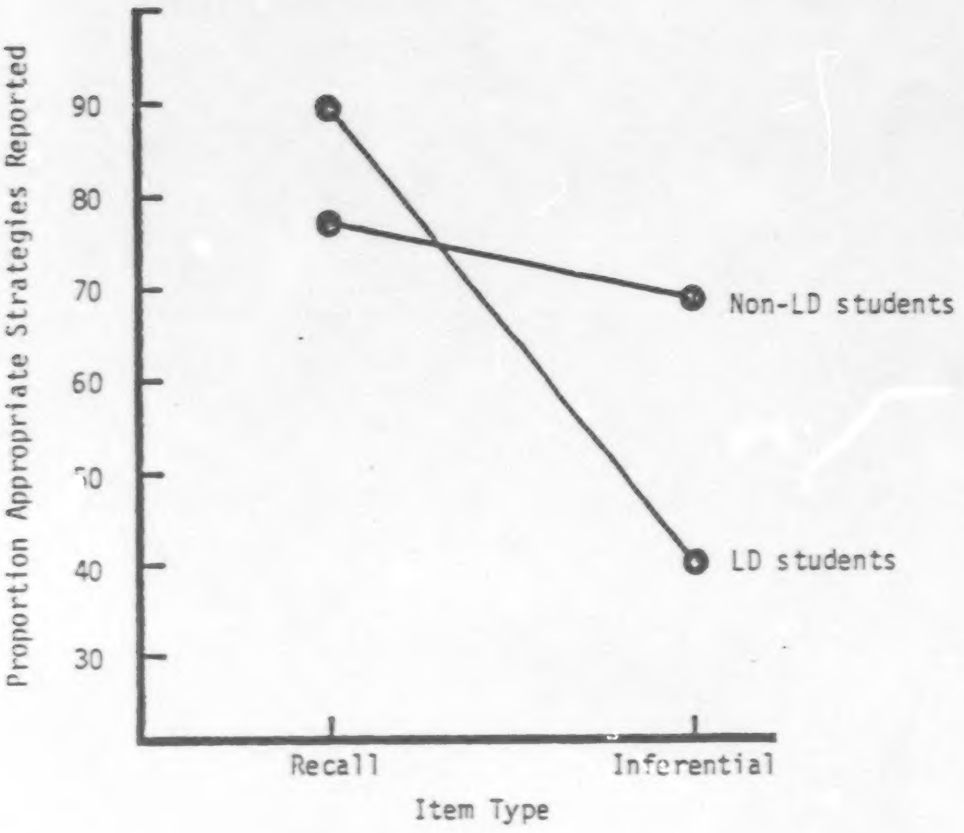
Figure 2. Proportion correct by strategy used on inferential items.

Figure 3. Illustration of two-way interaction of group by reported use of appropriate strategy.



*Fewer than 25% of students in this group reported Inappropriate or Reasoning strategies.


100
90
80
70
60
50
40
30
20
10
Percent Correct
Non-LD students
LD students
Inappropriate
Referring
Reasoning
Strategy


Proportion Appropriate Strategies Reported
90
80
70
60
50
40
30
Non-LD students
LD students
Recall
Inferential
Item Type

Spontaneously Employed Test-Taking Strategies of High and Low Comprehending Elementary School Children

Thomas E. Scruggs, Karla Bennion, and Steve Lifson
Exceptional Child Center
Utah State University

If important decisions are to be based on the results of standardized reading tests, student scores should provide the best possible estimate of reading performance. Unfortunately, the results of past research have indicated that student standardized test performance can be influenced by factors other than knowledge of test content. One of these factors, test-wiseness, includes time using strategies, error avoidance strategies, guessing strategies, and deductive reasoning strategies.

A question emerges concerning the extent to which elementary school students employ "test-taking" strategies when faced with difficult or ambiguous reading test items. Do students spontaneously use such strategies (that is, without being trained)? If so, which strategies (if any) are effective in obtaining correct answers?

To address those questions in the present study, the reading test performance of elementary school children was examined. In Experiment 1, two areas were investigated: (a) the strategies spontaneously employed by students to answer reading test items, and (b) the relative effectiveness of these strategies in increasing reading test scores. Experiment 2 examined the possible difference in use or utility of these strategies between average and learning disabled (LD) third graders.

EXPERIMENT 1

Method

A sample reading test based upon items from the Stanford Achievement Test (SAT) was administered to 31 elementary age Caucasian students (15 girls, 16 boys) attending summer classes in a western rural area. Students were selected so that a range of abilities as well as grade levels (1-6) were represented.

All students were seen individually by one of four examiners. Students were given selections from the SAT taken from the level one year higher than their assessed grade level on the Woodcock Reading Achievement Test, Passage Comprehension subtest. In this manner, a similar difficulty level was provided for each student. Most students were able to answer correctly approximately two-thirds of the test questions.

Students were then told to read aloud each test question (as well as the reading passages in the reading comprehension subtest), and to read aloud whichever of the distractors they chose to read. They were neither encouraged nor discouraged from reading each distractor. As soon as students had answered a test question, they were asked to rate their level of confidence in their response. After students had finished each subtest, they were asked to re-read the questions and tell the examiner why they had chosen the answer they did. The examiner recorded reading errors, confidence levels, attention to distractors, reference to reading passage, and reported strategies.

Results

It was found that all strategy responses could be classified within a 10-level hierarchy which strongly predicted probability of correct responding. Proportion of correct responses was computed across subjects for each type of strategy used and are shown in Figure 1.

These strategies were collapsed into five logical categories (skipping, procedural error, guessing strategy, deliberate strategy, and "knowing") and computed point biserial correlations for each subject. The median correlation between item score and reported strategy was .54 ($\underline{p} < .01$). No differential effects were seen by age or ability level, possibly due to the diverse nature of the sample.

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April, 1984. Presenter: Thomas E. Scruggs, Utah State University, UMC 68, Logan, UT 84322.

BEST COPY AVAILABLE

BEST COPY AVAILABLE

Students had a reasonably good idea of whether they had answered a test question correctly or not. When students reported being "very sure" their answer was correct, they were in fact correct 81% of the time. When they reported being "somewhat sure," they were correct only 13% of the time, and when they reported being "not sure", they obtained correct answers in only 7% of the cases. These figures are somewhat misleading, however. If looked at another way, the results seem different: when students answered correctly, they also reported being "very sure" the answer was correct in 56% of the cases.

A great deal of carelessness was observed in attention to all distractors. When students answered incorrectly, in 40% of the 302 cases they had not read all distractors. When students answered questions correctly, they had attended to all distractors in 73% of the 577 cases.

The results of Experiment 1 provided valuable general information about the manner in which children respond to reading achievement test items. However, the diversity of the population, in age and ability level, was thought to have obscured direct investigation of specific differences with respect to specific ability levels. Experiment 2, therefore, was conducted in order to determine whether differences in strategy use existed between a sample of learning disabled children and a sample of children not so classified. In order to clarify interpretation, grade level was held constant, and the number of subtests was reduced to two.

EXPERIMENT 2

Method

Subjects were 32 third grade students attending public schools in a rural area of a western state. Twelve were classified as learning disabled (LD) according to local school district criteria and P.L. 94-142, and 20 were regular classroom students, none of whom were referred for special services and who were thought by their teachers as functioning within a normal range of performance. Sex was evenly represented in LD and non-LD groups.

Two reading tests were constructed from items taken from the Stanford Achievement Test. Items were drawn from the Primary 2 battery for the instrument used with the LD group, and the Intermediate 1 level served as the source for the regular classroom group. The tests each contained three passages with 15 dependent questions. Items were adjusted to ensure that 14 questions (10 content, 4 inference) remained on each form. Comprehension questions were left in their original order in relation to the passage. The questions were renumbered to avoid gaps where passages did not follow each other sequentially in the original test. In addition, three items from the letter-sound test (level P3) were selected. These consisted of a stimulus word with a letter or letters underlined representing a sound that the student had to identify in the three options given below the stimulus word. These items were selected to include a distractor that closely matched the initial consonants of the stimulus word. For example, in the item:

blind
0 blink
0 nibble
0 leaned

"leaned" is the correct answer, since it contains the same sound as the underlined nd in the stem, and "blink" is the inappropriate distractor, since it contains the same initial consonant blend.

Subjects were seen individually by one of two examiners. They were asked to read the passages and questions aloud and mark answers they thought were correct. They were then told that they would be asked if they were sure/not sure that the answer they had given was correct, and the manner in which they had chosen the particular answer. The subject's response to the questions, "How did you choose that answer?" and "Are you sure or not sure of your answer?" were recorded verbatim on the protocol. Words the experimenters had deemed essential to answering the questions (key words) were marked in the examiner's copy of the instrument, and errors in these words were noted as the child read aloud.

Results

Test items were scored for correctness, confidence in answer (sure/not sure), and type of strategy reported. Two students from the non-LD group, who had misread more than 25% of the keywords, were excluded from further analysis. The responses given by the subjects were divided into seven logical categories:

> 1 = Don't know; 2 = Guessed; 3 = External source of knowledge (e.g., "I know all fish have scales"); 4 = Refers to passage (e.g., "I read it"); 5 = Quotes directly (e.g., "It says here that . . ."); 6 = Eliminates options known to be incorrect; 7 = Other reasoning (e.g., "It said comforted in the story. That sort of means relieved.)

BEST COPY AVAILABLE

each response was then evaluated in terms of those seven categories. Percent of agreement for scoring was assessed at 100% after each examiner scored 25% of the other examiner's protocols.

Proportion correct by collapsed strategy group (inappropriate = strategies 1-3; referring = strategies 4-5; reasoning = strategies 6-7) was computed for item type and student group and is given in Figures 2 and 3. As can be seen, a monotonically increasing trend is seen for both groups.

A $\underline{t}$ test applied to percent of keywords read incorrectly indicated that the groups did not differ significantly with respect to reading difficulty, $\underline{t}(29) = .37$, $\underline{p} > .20$. Overall, LD students misread 6.6% of total keywords and non-LD students misread 6.75% of keywords.

Reported strategy data were scored for appropriateness of reported strategy. Strategies were considered appropriate if students reported referring to the passage on a recall question (strategy 4 or 5), or if they reported a reasoning strategy in response to an inferential question (strategy 6 or 7). Proportion of appropriate responses were then entered into a 2 group (LD vs. non-LD) by 2 item type (direct recall or inferential) analysis of variance (ANOVA) with repeated measures on the item type variable. Because of the unequal group frequencies, a least-squares method of analysis (Winer, 1971) was employed. Significant differences were found for item type, $\underline{F}(1,29) = 9.19$, $\underline{p} < .01$, and for interaction, $\underline{F}(1,27) = 7.58$, $\underline{p} < .05$. Figure 4 depicts graphically the interaction effect. Although both LD and non-LD students reported a high proportion of "referring to text" strategies (89% vs. 77%, respectively), large differences emerged in proportion of reasoning strategies applied to inferential questions (39% vs. 70%, respectively). Nonsignificant differences were observed for overall group means, $\underline{F}(1,29) = 1.54$, ns.

Analysis of confidence reports indicate that both groups were similar with respect to reported level of confidence on "referring to passage" strategies with LD students reporting confidence in 85% of the cases and non-LD students reporting confidence in 92% of the cases. These reports were similar to actual performance, with correct scores of 81% and 86% on these items for LD and non-LD groups, respectively. On reasoning strategies, however, a much different picture emerged. Average students were correct on 83% of inferential items, but reported confidence on an average of 71% of the items. The LD students, on the other hand, reported being confident on an average of 95% of the cases, but were in fact correct in only 63% of these cases.

Items on the letter-sound subtest were scored for responses which suggested attention to an inappropriate distractor. This inappropriate distractor took the form of an initial consonant blend present in the stem, but not underlined. Number of inappropriate distractors chosen was compared by group, and differences found to be significant, $\underline{t}(28) = 2.47$, $\underline{p} < .05$. The LD children chose the inappropriate distractor in 52% of the cases, while the non-LD children chose the inappropriate distractor in only 24% of the cases.

Discussion

It has been seen that the present sample of LD third graders, with reading ability controlled for, differed from their more average counterparts with respect to (a) proportion of appropriate reasoning strategies reported for inferential comprehension questions, (b) performance and confidence level for items in which reasoning strategies had been reported, and (c) choice of an inappropriate distractor on a letter-sounds test. On the other hand, LD students did not differ from their more average counterparts with respect to appropriate strategy use on recall items. Generally, this sample of LD children was seen to report fewer reasoning strategies, when appropriate, on reading comprehension test items than did their more average counterparts, and to be less successful on those items for which they did report reasoning strategies. The inappropriately high levels of confidence exhibited by the LD students on items for which reasoning strategies had been applied is supportive of a theory of a developmental deficit in "meta-cognitive abilities" (cf. Torgesen, 1977), in that inappropriately high levels of confidence in task performance are often seen in younger children.

Reading comprehension, clearly, is a construct which seems to resist precise analysis and for which many theoretical orientations exist. If one does look at recall and inference as two component parts of reading comprehension, however, it appears from the present investigation that relative strategy and performance deficits exist on the part of LD children on inference questions, but not on recall questions, with reading ability controlled for.

Replication is necessary to further support and refine these findings. The present results suggest that LD children may benefit from specific training in (a) identifying inference questions, (b) selecting appropriate strategies relevant to those questions, and (c) successfully applying such strategies to reading content.

Teaching Test-Taking Skills to Elementary Grade Students: A Meta-Analysis

Thomas E. Scruggs, Karla Bennion, and Karl White
Exceptional Child Center
Utah State University

Running head: IMPROVING ACHIEVEMENT TEST SCORES

Abstract

Results of 24 studies which investigated the effects of training elementary school children in test-taking skills on standardized achievement tests were analyzed using meta-analysis techniques. In contrast to all previous reviewers, the results of this analysis suggest that training in test-taking skills has only a very small effect on students' scores on standardized achievement tests. Longer training programs are more effective, particularly for students in grades 1-3, and for students from low socioeconomic status background. Results from previous reviews of this body of literature are critiqued and explanations offered as to why the results of the present investigation are somewhat contradictory to previous reviewers' conclusions. Suggestions for further research are given.

Teaching Test-Taking Skills to Elementary Grade Students: A Meta-Analysis

Since the seminal work of Millman, Bishop, and Ebel (1965), much attention has been directed to the influence of test-taking skills, or "test-wiseness," on scores of achievement tests. Assumptions from the past have included that test-wiseness is a substantially separate variable not strongly correlated with intelligence (Diamond & Evans, 1972), that test-taking skills are alterable by training, and that these skills would transfer to higher scores on achievement tests (Ford, 1973; Fueyo, 1977; Sarnacki, 1979).

Training materials have been created (some of which are commercially available) to teach "test-taking skills" (e.g., Mini-Tests, 1979 and Test-Taking Skills Kit, 1980), and claims have been made that such training leads to increased test scores (e.g., Fueyo, 1977; Jones & Ligon, 1981; Samson, 1984). The rationale for such training programs stems from the common practice of utilizing results from achievement tests to assist in making decisions about educational placement, programming, and evaluation. To the degree that achievement tests are measuring test-taking skills rather than mastery of the content being tested (e.g., reading, math), decisions about placement, programming, and evaluation may be incorrect (see Ebel, 1965, for additional discussion). Promoters of teaching test-taking skills have claimed that students would obtain higher scores if deficiencies in test-taking skills were remediated, thus resulting in a more valid indicator of how well the student had mastered the content the test was designed to assess.

Although efforts to reduce measurement error in standardized achievement testing are commendable, several questions remain:

1. Although many people have concluded that test-taking skills training leads to increased test scores, is that position consistently supported empirically, and what is the magnitude of typically obtained effects?

2. Can the cost of typical test-taking training programs be justified in view of the magnitude of observed effects and the alternative uses of the same resource (i.e., is it cost-effective?)?

3. Are some types of training more valuable than others in increasing performance on achievement tests, and are some groups of children more likely than others to benefit from such training? The purpose of the present investigation was to integrate the results from previous research to answer the preceding questions as they pertain to standardized achievement tests with elementary school-aged children.

Review of Previous Work

Several reviewers have previously examined the effects of teaching test-taking skills (Bangert-Drowns, Kulik & Kulik, 1983;

Ford, 1973; Fueyo, 1977; Jones & Ligon, 1981; Sarnacki, 1979; Taylor, 1981). A summary of the characteristics and conclusions of these reviewers is shown in Table 1.

Insert Table 1 about here

All previous reviewers concluded that test-taking skills could be taught effectively and resulted in benefits for children (including higher achievement test scores). Unfortunately, except for Bangert-Drowns et al. (1983) and Taylor (1981), previous reviews failed to indicate the procedures or criteria for including research studies in their review, did not cite and critique prior reviews, and apparently only analyzed results of the primary research included in their review in terms of the original researcher's conclusions. As will be shown below, all of the reviewers failed to include a substantial number of studies with elementary aged children. Consequently, one cannot be confident that results cited in these reviews are representative of available reseach. It is also difficult to draw conclusions about the magnitude of the alleged effect of training students in test-taking skills since most of the reviewers stated only that differences were found, or improvement was noted, and occasionally referred to statistically significant differences between groups. Without knowing more about the magnitude of the effect attributable to teaching test-taking skills, it is difficult to draw conclusions about whether it is likely to be a wise investment to divert resources from other activities (e.g., teaching reading) to teach test-taking skills.

Taylor (1981) conducted an excellent review on the effects of practice, coaching, and reinforcement on test scores. This investigation focused upon all age levels and on group-administered as well as individually administered tests. The great majority of studies selected, in fact, concentrated on either IQ tests or non-elementary age populations; consequently, a substantial number of studies which investigated the effects of training achievement test-taking skills with elementary-aged children were not included in her review.

The most comprehensive analysis to date of the effect of teaching test-taking skills on achievement test scores was a meta-analysis recently completed by Bangert-Drowns et al. (1983). The effect of teaching test-taking skills for elementary-and secondary-aged children was analyzed by computing a standardized mean difference effect size for each study (Glass, 1977) to indicate the extent to which achievement test scores were altered by training. This was a substantial improvement from most earlier reviews which relied primarily on authors' conclusions or tests of statistical significance without indicating the magnitude of effects. Knowing the magnitude of improvement is very important so that practitioners can make judgments concerning whether the investment in training is cost-effective compared to what else could have been accomplished with that time. Bangert-Drowns et al. (1983) concluded that teaching test-taking skills raised standardized achievement test scores by .25 standard deviations--enough to raise the typical student from the 50th to the 60th percentile. They also concluded that length of training program was positively related to effect size; drill and practice was less effective than training in "broad cognitive skills;" and effectiveness of training was not affected by identifiable subject characteristics or other characteristics of the program.

Although Bangert-Drowns et al. provided valuable information, their study is limited by several factors. First, a number of studies have been done which were not included in their review. Secondly, although indicators of study quality were coded, there was no report of efforts to determine if there were differential effects for studies of high versus low quality. It may be, for example, that investigations of lower quality produce effect sizes which are substantially different (and also less credible) than studies of high quality.

Third, their decision to average all outcomes from a given study into one measure of effect size can be misleading. For example, Levine (1980) randomly assigned low SES and not low SES fifth graders to either test-taking training or control groups and collected data on students' scores on standardized reading achievement and an assessment of "test-wiseness". Four obvious effect sizes are possible: low SES experimental versus control for reading and test-wiseness; and not low SES experimental versus control for reading and test-wiseness. These four effect sizes range from .38 to 1.52 and average .90. To report only the average of all four is not only misleading, but irretrievably obscures important differences between types of subjects and types of outcome (e.g., in this study the effects for low SES subjects were much larger than "not low SES" subjects for both outcomes, and effects for test-wiseness were much larger than reading achievement for both groups).

Finally, in some instances Bangert-Drowns et al. appear to have used inappropriate computations for determining the effect size. For example, in the Romberg (1978) study, classrooms were randomly assigned to treatments, and class averages were used as the unit of analysis. While the use of classroom means as the unit of analysis is an appropriate statistical procedure (Peckham, Glass, & Hopkins, 1969), the standard deviation of group means will generally be much smaller than the within-group standard deviation. The use of the between-class standard deviation will result in a much larger effect size and will not be comparable to studies for which the within-group standard deviation was used. In the Romberg study, Bangert-Drowns et al. apparently used the between-class standard deviation for achievement test scores and obtained an effect size of .48. By contrast, the present authors estimated the effect size (since within-group standard deviations were not reported) by converting the reported percentile scores to Z scores and using differences in Z scores as the effect size. This procedure yielded an effect size based on the within-group standard deviation of only .14--less than one third the magnitude of Bangert-Drowns et al. estimate.

Other important questions remain unaddressed by Bangert-Drowns et al. (1983). First, many investigations believe that the training of test-taking skills is particularly beneficial for children in low socioeconomic settings (e.g., Jones & Ligon, 1981; Jongsma & Warshauer, 1975). Thus, it is important to determine whether teaching test-taking skills has a differential effect on children of low socioeconomic status than it does on children who do not come from such groups. Secondly, it is important to determine whether the effects of training in test-taking skills are different for children of different ages. In the Bangert-Drowns et al. study, students in grades 1 to 6 were combined into one category. Third, it is important to replicate their findings about length of training and type of training, and to determine whether there are any other important concomitant variables or interactions among variables not identified by Bangert-Drowns et al. Finally, it is important to know whether studies of adequate validity produce different effect sizes from studies of less than adequate validity, and whether there is a differential effect for different types of dependent measures (e.g., achievement tests, measures of test-wiseness, student attitude).

Procedure

Location of studies. Several procedures were used to find as many studies as possible which investigated the effect on group-adminstered standardized achievement test scores of teaching test-taking skills to elementary-aged school children. Studies which examined attempts to improve, for example, scores on individualized achievement tests or IQ tests were excluded from this analysis. Also excluded from analysis were studies which investigated the effects of training on achievement test performance of students of greater than 6th grade level.

Studies were located by first conducting a computer-assisted search of Dissertation Abstracts International, Psychological Abstracts, and Educational Resources Information Center (ERIC) data bases. Studies found in this way were examined to determine whether they contained references to other appropriate studies. Previous reviews of research on teaching test-taking skills (Bangert-Drowns et al., 1983; Ford, 1973; Fueyo, 1977; Jones & Ligon, 1981; Sarnacki, 1979; Taylor, 1981) were also examined for additional studies. Twenty-four experimental studies of the effects of teaching test-taking skills on achievement tests for students in grades 1 through 6 were located. This number is 70% greater than the greatest number of studies involving achievement tests for elementary school children found by any previous reviewer.

Coding. Each study was coded for 14 different variables which described the type of subjects with whom the research was conducted, the type of training provided, the experimental design used, and the type of outcome data collected. The specific variables coded are reported in Table 2 in the results section. Interrater consistency was established by having two independent reviewers code each article. Wherever disagreement occurred, differences were resolved by discussion.

To enable the comparison of all outcomes across all studies, an effect size for each relevant comparison was computed (Glass, McGaw, & Smith, 1981). Effect size was defined as the mean difference between two groups divided by the standard deviation of the control group. When means and standard deviations are not reported in a study, effect sizes can also be calculated from other statistics such as t and F. Basic conventions for determining which effect sizes to code, and methods of calculation when means and standard deviations were not available, are given in Casto, White, and Taylor (1983).

In addition, obtained effect sizes were adjusted using Hedges' (1981) formula for bias correction of the effect size estimator before analyses were done. Although the correction procedure was used for all results in the present study, the authors agree with Bangert-Drowns et al. that the overall difference in effect sizes due to this correction procedure was trivial (only 1 out of 65 effect sizes changed by more than .01 of an effect size).

Results and Discussion

The 24 investigations of the effect of teaching test-taking skills resulted in 65 effect sizes which were relatively evenly distributed among studies. The mean effect size for all comparisons including achievement tests, tests of test-wiseness, self-esteem, and anxiety, was .21, a figure which is consistent with that of Bangert-Drowns et al. but should be interpreted with caution since it is the average across different types of dependent measures, studies of differing quality, and students with different characteristics.

Table 2 shows the mean effect size for all levels of the different variables coded in the meta-analysis. As can be seen,

Insert Table 2 about here

the average effect size for studies with adequate validity is relatively close to that of studies with inadequate validity (.20 vs. .29). Although this suggests that it may not be necessary to account for quality of study in interpreting the impact of training students in test-taking skills, further examination of Table 2 shows that this is not the case. In particular, we note that the average of 44 effect sizes for achievement test scores from studies of adequate validity is .10, while the average of 6 effect sizes from adequate studies measuring "test-wiseness" is .71--almost 10 times as large. There are also no measures of test-wiseness or measures such as anxiety, self-esteem, and attitude towards the test, which come from studies with inadequate validity. Thus, the apparent equivalence in average effect sizes between studies of adequate validity and inadequate validity is largely attributable to the fact that outcomes other than achievement all come from studies of adequate validity and yield substantially higher effect sizes than measures of achievement.

The mean effect size for achievement test scores from studies of adequate validity is only .10 compared to an average of .29 for achievement test scores for studies with inadequate validity. This contrasts sharply with the findings of Bangert-Drowns et al. who reported an average effect size of .25. Part of the reason that Bangert-Drowns et al. found a higher average effect size may have been that they collapsed several different outcome measures from the study into one average effect size. As noted above, this can be misleading and prevents analyses of important issues.

Because there is such a dramatic difference in average effect size between studies with adequate validity and studies with inadequate validity, and between measures of achievement and other measures, the remaining analyses will focus primarily on effect sizes of achievement tests from studies with adequate validity.

The mean effect sizes for achievement test scores from studies with adequate validity for different levels of length of treatment, SES level, and grade level are shown in Table 3.

Insert Table 3 about here

As can be seen, there was considerable difference between interventions which were less than 4 hours and those which were 4 or more hours (.04 vs. .29). A similar finding was seen when results of achievement test scores were broken down by grade level. When treatments were administered to students in the primary grades (1-3), the average effect size on standardized achievement tests was only .01. From grades 4-6, however, the mean effect size for achievement tests was much higher, .20. The difference between students of differing socioeconomic background was very slight ($\overline{ES}$ = .14 vs. $\overline{ES}$ = .09), with a very small advantage for students from low socioeconomic backgrounds.

Even more interesting than the average effect size for different levels of these three variables are the interactions between the variables. As can be seen in Figure 1, for treatments involving less than 4 hours, students in the primary grades exhibited slightly negative effect sizes ($\overline{ES} = -.12$) while students from grades 4 through 6 had an average effect size of .19. For students receiving more than 4 hours of training, however, there is no difference--students in both grades 1-3 and 4-6 had an average effect size of .29. Although the mean effect size for students in grade 1-3 with 4 or more hours of treatment is based on only four studies, these data are provocative and require further investigation. More specifically, it appears that for older students, a short amount of training in test-taking skills may result in substantial improvement. However, for younger children, it takes much more training before there are observable benefits.

Figure 2 shows another interesting interaction between length of training and socioeconomic status. With less than 4 hours of treatment, neither "low SES" nor "not low SES" subjects benefited appreciably (average effect sizes are .05 and .08). With high levels of treatment, students from low socioeconomic backgrounds benefit more than twice as much as students who are not from low socioeconomic backgrounds (average effect size = .44 vs. .20). Again, this finding requires further replication before confident conclusions can be drawn, but it suggests that authors who have contended that training in test-taking skills is most important for students from low socioeconomic background (e.g., Jones & Ligon, 1981; Jongsma & Warshauer, 1975) may be correct.

Before drawing conclusions about the efficacy of training students in test-taking skills, it is important to comment briefly on the differences in average effect sizes between outcomes of achievement test scores ($\overline{ES}$ = .10), tests of test-wiseness ($\overline{ES}$ = .71), and measures of anxiety, self-esteem, and attitude towards tests ($\overline{ES}$ = .44). Admittedly, the measures other than scores on achievement tests are based on a very limited number of studies, so one should be cautious in drawing conclusions. However, from these data, it appears that tests of test-wiseness are more sensitive to training effects. One explanation for this much larger average effect size is that the training program is "teaching to the test." The fact that high scores on tests of test-wiseness are not necessarily related to higher achievement test scores suggests that the relation between test-wiseness and high scores on achievement tests is not very strong. It should be remembered that the primary argument for providing training in test-taking skills to students has always been related to the need to reduce measurement errors in the child's standardized test score. To the degree that that is happening, it has been assumed that test scores would go up. Although the fact that test scores are not going up appreciably is not proof that scores are not more accurate, it still leaves the burden of proof upon those who claim that training in test-taking skills is beneficial. Higher scores on tests of test-wiseness are not sufficient evidence for those benefits.

Conclusions

As noted earlier, this integrative review was designed to answer the following three questions:

1. To what degree is the popular position that training in test-taking skills is beneficial for children supported by empirical evidence?

2. Do the data about the effect of teaching test-taking skills justify the use of resources for this purpose as opposed to alternative uses of the same resource?

3. Are some types of training more effective or are some groups of children more likely to benefit from training in test-taking skills?

In response to the first question, the results of this review stand out in contrast with all previous reviewers of the effects of training in test-taking skills. The most credible evidence (results from high quality studies limited to scores on standardized achievement tests), at least as it pertains to elementary school-aged children, does not demonstrate a sizeable benefit for teaching test-taking skills. The reason for these different conclusions is partly attributable to the use of more systematic techniques than used by many of the previous reviewers to identify the magnitude of the effect and how that effect covaried with other variables. More importantly, a larger number of studies was identified and quality of study and type of outcome was accounted for.

Is training in test-taking skills cost effective? The answer is not clear-cut. Clearly, benefits of a tenth of a standard deviation are relatively small (less than one month worth of gain in reading for an average third grader), but they were obtained at relatively little cost. Even the longest training program lasted only 20 hours, and the majority of effect sizes came from studies in which training lasted less than 4 hours. The question also depends in part on whether one is talking about children in grades 1-3 or grades 4-6. These data suggest that for older children, a limited amount of training can have a discernible effect. For younger children, more training is necessary. Also, the fact that a few studies (unfortunately, it is a very limited number) suggest that training in test-taking skills has some positive impact on anxiety, self-esteem, and attitude towards tests should not be forgotten. However, before it is accepted as fact, more research needs to be done. It is clear that a comprehensive analysis of previous research on training test-taking skills suggests that the benefits are not nearly so great as has typically been concluded.

Data from the meta-analysis do suggest that training in test-taking skills is differentially effective for various subgroups of children. The interactions between length of treatment and grade level, and length of treatment and SES are particularly provocative and deserve further research. In general, the meta-analysis supports the conclusion of Bangert-Drowns et al. that longer training programs are more effective. As a general strategy, it also appears that training is more effective in the upper elementary grades than in the lower elementary grades. Whether or not a training package includes practice tests, reinforcement, or drill and practice does not seem to be an issue about which we have sufficient data to draw conclusions. More research is needed before we can decide what types of training are most effective.

Should training in test-taking skills be pursued? Hopefully, the results of this analysis will temper some of the unfounded enthusiasm in support of training children in test-taking skills. However, it would be unwise to conclude that training in test-taking skills is unwarranted or detrimental. Although the effects of such training are small, the investment is relatively cheap, and there is some evidence that for particular groups of children, training in test-taking skills can have substantial effects. Those tentative conclusions need further research, but indicate an area worth pursuing.

References

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. _Review of Educational Research_, _53_, 571-585.

Casto, G., White, K. R., & Taylor, C. (1983). An Early Intervention Research Institute: Efficacy and cost studies in early intervention. _Journal of the Division for Early Childhood_, _7_, 5-17.

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of test-wiseness. _Journal of Educational Measurement_, _9_(2), 145-150.

Ebel, R. L. (1965). _Measuring educational achievement_. Englewood Cliffs, NJ: Prentice-Hall.

Ford, V. A. (1973). _Everything you wanted to know about test-wiseness_. Princeton, NJ: Educational Testing Service. (ERIC Reproduction Service No. ED 093 912)

Fueyo, V., (1977). Training test-taking skills: A critical analysis. _Psychology in the Schools_, _14_, 180-184.

Glass, G. V (1977). Primary, secondary, and meta-analysis of research. _Educational Research_, _5_, 3-8.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). _Integrating research studies: Meta-analysis of social research_. Beverly Hills, Ca.: Sage Publications.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 5, 107-128.

Jones, P., & Ligon, G. D. (1981). Preparing students for standardized testing: A literature review. Austin, Tx: Austin Independent School District. (ERIC Document Reproduction Service No. 213 768)

Jongsma, E. A., & Warshauer, E. (1975). The effects of instruction in test-taking skills upon student performance on standardized achievement tests. New Orleans, LA: New Orleans University, Department of Elementary and Secondary Education. (ERIC Document Reproduction Service No. ED 114 408)

Levine, M. A. (1980). Training in test-wiseness on reading scores of low and middle SES pupils (Doctoral dissertation, Yeshiva University, 1979). Dissertation Abstracts International, 40, 6242A. (University Microfilms No. 80-12,678).

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test wiseness. Educational and Psychological Measurement, 25, 707-726.

Mini-tests (1979). New York: Educational Solutions, Inc.

Peckham, P. D., Glass, G. V, & Hopkins, K. D. (1969). The experimental unit in statistical analysis. The Journal of Special Education, 3(4).

Romberg, E. (1978). The effects of test-taking skills and attitudes on validity of standardized achievement test scores of inner-city children (Doctoral dissertation, University of Maryland, 1977). Dissertation Abstracts International, 39, 832A. (University Microfilms No. 73-12,646).

Samson, G. E. (1984). Effects of training in test-taking skills on achievement: A quantitative analysis. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 49(2), 252-279.

Taylor, C. (1981). The effect of reinforcement and training on group standardized test behavior. Unpublished doctoral dissertation, Utah State University, Logan.

Test-taking skills kit (1980). Herndon, VA: Evaluation and Assessment Service, Inc.

META-ANALYSIS REFERENCES

Butler, D. D. (1983). An assessment of the effects of instruction and practice on the test-wiseness of fourth graders as measured by changes in standardized test scores (Doctoral dissertation, University of Wyoming, 1982). Dissertation Abstracts International, 43 (7-A), 2322.

Callenbach, C. (1977). The effects of instruction and practice in content-independent test-taking techniques upon the standardized reading test scores of selected second-grade students. Journal of Educational Measurement, 14, 335-341.

Costar, E. (1980). Scoring high in reading: The effectiveness of teaching achievement test-taking behaviors. Elementary School Guidance and Counseling, 15, 157-159.

Crowe, D. E. (1982). The use of practice programs to improve test scores of elementary school students (Doctoral dissertation, University of South Carolina, 1981). Dissertation Abstracts International, 42 (7-A), 3116.

Derby, T. L. (1979). The effects of instruction in selected aspects of test-wiseness on the administration of standardized reading items in the upper elementary school (Doctoral dissertation, University of Pennsylvania, 1979). Dissertation Abstracts International, 39 (12-A), 7236.

Dillard, M., Warrior-Benjamin, J., & Perrin, D. W. (1977). Efficacy of test-wiseness on test anxiety and reading achievement among Black youth. Psychological Reports, 41, 1135-1140.

Eakins, D. J., Green, D. S., & Bushnell, D. (1976). The effects of an instructional test-taking unit on achievement test scores. Journal of Educational Research, 70, 67-71.

Jongsma, E. A., & Warshauer, E. (1975). The effects of instruction in test-taking skills upon student performance on standardized achievement tests. New Orleans, LA: New Orleans University, Department of Elementary and Secondary Education. (ERIC Document Reproduction Service No. ED 114 408).

Kalechstein, P., Kalechstein, M., & Docter, R. (1981). The effects of instruction on test-taking skills in second-grade black children. Measurement and Evaluation in Guidance, 13(4), 198-202.

Lagana, J. L. (1979). The effects of incentive motivation and test-wiseness coaching on the standardized reading test scores of third-grade students (Doctoral dissertation, University of California at Los Angeles, 1978). Dissertation Abstracts International, 39, 4198-4199.

Levine, M. A. (1980). Training in test-wiseness on reading scores of low and middle SES pupils (Doctoral dissertation, Yeshiva University, 1979). Dissertation Abstracts International, 40, 6242A. (University Microfilms No. 80-12,678).

Luddeke, N. S. (1972). The effect of motivational programs on standardized achievement test performance of disadvantaged third graders at two levels of test difficulty (Doctoral dissertation, University of Cincinnati, 1971). Dissertation Abstracts International, 33 (6-A), 2820.

Romberg, E. (1978). The effects of test-taking skills and attitudes on validity of standardized achievement test scores of inner-city children (Doctoral dissertation, University of Maryland, 1977). Dissertation Abstracts International, 39, 832A. (University Microfilms No. 73-12,646).

Schuller, S. M. (1979). A large-scale assessment of an instructional program to improve test-wiseness in elementary school students. New York: Educational Solutions, Inc. (ERIC Document Reproduction Service No. ED 189 143).

Shisler, C. L. (1973). A study of test performance of first graders under three conditions of motivation (Doctoral dissertation, University of South Carolina, 1973). Dissertation Abstracts International, 1714-A.

Slaughter, B. A. (1976). An examination of the effects of teaching and practice in test-taking skills on student performance on a standardized achievement test (Doctoral dissertation, University of Pittsburgh, 1976). Dissertation Abstracts International, 37, 1505A. (University Microfilms No. 76-19,931).

Stephenson, P. C. (1976). Improving the learning disabled child's score on machine-scored tests. Journal of Learning Disabilities, 9(2), 17-19.

Suber, J. S. (1980). The effects of a practice test and days for administration on the demonstrated achievement level of low achieving third grade students in South Carolina (Doctoral dissertation, University of South Carolina, 1979). Dissertation Abstracts International, 40 (11-A), 5833.

Taylor, C. E., & White, K. R. (1983). The effect of reinforcement and training on group standardized test behavior. Journal of Educational Measurement, 19(3), 199-209.

Thomas, R. J. (1977). The effects on three methods of test anxiety and the achievement test performance of elementary students: Providing test-taking information, test-wiseness training, and systematic desensitization (Doctoral dissertation, University of Wisconsin, Madison, 1976). Dissertation Abstracts International, 37 (9-A), 5717-5718.

Tinney, R. E. (1969). The effect of training in test-taking skills on the reading test scores of fifth grade children of high and low socioeconomic levels (Doctoral dissertation, University of Minnesota, 1968). Dissertation Abstracts International, 30, 595A. (University Microfilms No. 69-11,505).

Van Hoose, W. (1969). The efficacy of counseling in the elementary school. Ohio State University. (ERIC Document Reproduction Service No. ED 033 394).

White, K. R., Taylor, C., Friedman, S., Bush, D. & Stewart, K. (1983). An evaluation of training in standardized achievement test taking and administration: Final report of the 1981-82 Utah State Refinements to the ESEA Title I Evaluation and Reporting System. Utah State University and Utah State Office of Education.

Yearby, M. E. (1976). The effects of instruction in test-taking skills on the standardized reading test scores for white and black third-grade children of high and low socioeconomic status (Doctoral dissertation, Indiana University, 1975). Dissertation Abstracts International, 36, 4426A. (University Microfilms No. 75-23,438).

Table 1

Characteristics and Conclusions of Previous Reviewers of the Effect of Teaching Test-Taking Skills

Author/year	# of experimental studies cited	Methods for selecting studies specified?	Previous reviewers cited and critiqued	Outcomes of experimental studies cited in terms of	Conclusions about effectiveness of training test-taking skills	Variables cited which covary with effect of training	Type of studies included
Bangert-Drowns et al./1983	30	Yes	No	Standardized effect size	Effective $\overline{ES}$ = .25	Length of training program, type of training	Achievement tests; elementary and secondary level
Ford/1973	24	No	No	Conclusions	Effective	None	Achievement, IQ, and aptitude tests; preschool through adult
Fueyo/1977	19	No	No	Conclusions	Effective	None	Achievement, IQ, and aptitude tests; preschool through adult
Jones & Ligon/1981	5	No	No	Conclusions	Effective	Maintenance of effect Socioeconomic status	Achievement, IQ, and aptitude tests; preschool through adult
Sarnacki/1979	17	No	No	Conclusions	Effective	None	Achievement, IQ, and aptitude tests; preschool through adult
Taylor/1981	34	Yes	Yes	Standardized effect size	Effective $\overline{ES}$ = .62	Type of training, unit of administration, quality of study, type of test (achievement vs. IQ)	Achievement, IQ, and aptitude tests; preschool through adult

BEST COPY AVAILABLE

BEST COPY AVAILABLE

Table 2

Mean Effect Size for All Levels of All Coded Variables

		Adequate validity			Inadequate validity		
		$\overline{ES}$	SD_{ES}	N_{ES}	$\overline{ES}$	SD_{ES}	N_{ES}
All studies		.20	.40	55	.29	.33	10
Total sample size for study:	Small (0-75)	.32	.28	21	.40	.46	5
	Medium (76-150)	.11	.50	24			
	Large (150+)	.15	.30	10	.18	.08	5
Grade level:	1st-3rd	.03	.51	25	.14	.06	6
	4th-6th	.33	.39	30	.59	.54	3
Socioeconomic status level:	Low	.18	.37	37	.33	.36	8
	Not low	.24	.46	18	.11	.02	2
Use of reinforcement procedures as part of training:	No	.22	.40	48	-	-	-
	Yes	-.00	.43	7	.29	.33	10
Hours of training:	Less than 1 hr	.09	.43	14	.37	.47	5
	1 to 3 hrs	.09	.30	22	-	-	-
	4 hrs+	.40	.42	19	.20	.13	4
Use of practice tests as part of training:	No	.22	.43	42	.40	.46	5
	Yes	.12	.30	13	.16	.07	4
Ability level of students:	Mixed	.20	.52	47	.29	.33	10
	High ability	.09	.21	3	-	-	-
	Low ability	.31	.12	5	-	-	-
Type of assignment to groups:	Random	.27	.39	40	.30	.40	7
	Good matching	.24	.01	2			
	Poor matching	-.05	.37	13	.28	.10	3
Blinding of data collector:	Yes	.13	.44	34	.16	.07	4
	No	.31	.30	21	.38	.42	6
Type of outcome measure:							
Achievement test		.10	.33	44	.29	.33	10
Test-wiseness test		.71	.57	5	-	-	-
Other (anxiety, self-esteem, attitude)		.44	.36	6	-	-	-

$\overline{ES}$ = mean effect size for a particular group.

SD_{ES} = standard deviation of effect size distribution for a particular group.

N_{ES} = number of effect sizes on which a computation is based.

Note. Several other variables including Percent Male, Percent Handicapped, and Percent Minority were coded to determine whether mean effect size covaried with such subject characteristics. Results for those variables are not reported here because of infrequent reporting (e.g., Percent Handicapped could only be coded for 2% of the ES's), or lack of variance (e.g., 97% of the ES's for Percent Male fell between 47% and 54%).

Table 3

Mean Effect Sizes on Achievement Test Scores, Broken Down by Treatment Length, SES Level, and Grade Level

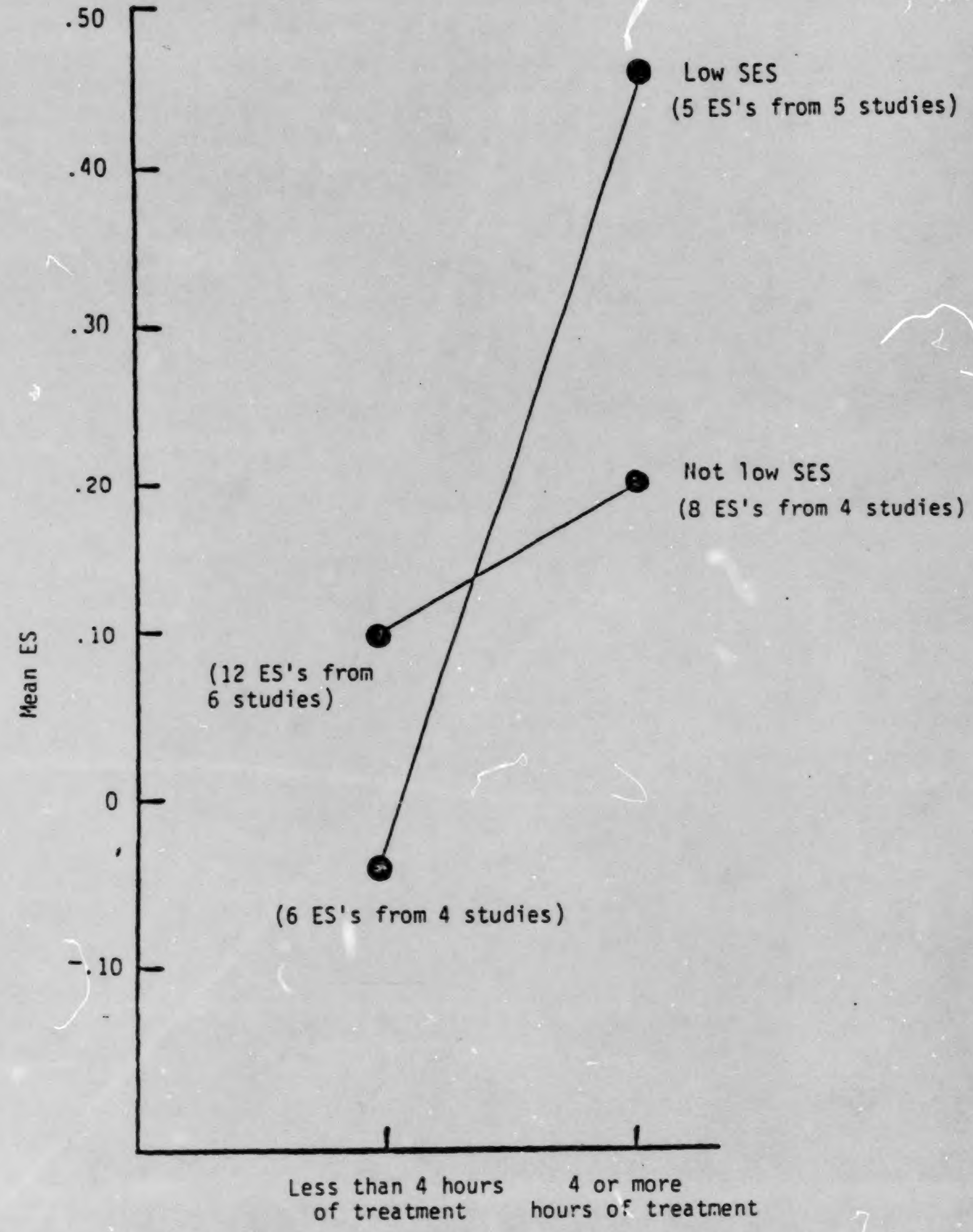
	Mean $\overline{ES}$	SD_{ES}	n_{ES}	$N_{Studies}$
Less than 4 hours of treatment	.04	.30	18	7
4 or more hours of treatment	.29	.31	13	8
Low SES	.14	.38	13	10
Not low SES	.09	.31	31	13
Grades 1-3	.01	.37	22	9
Grades 4-6	.20	.26	22	9

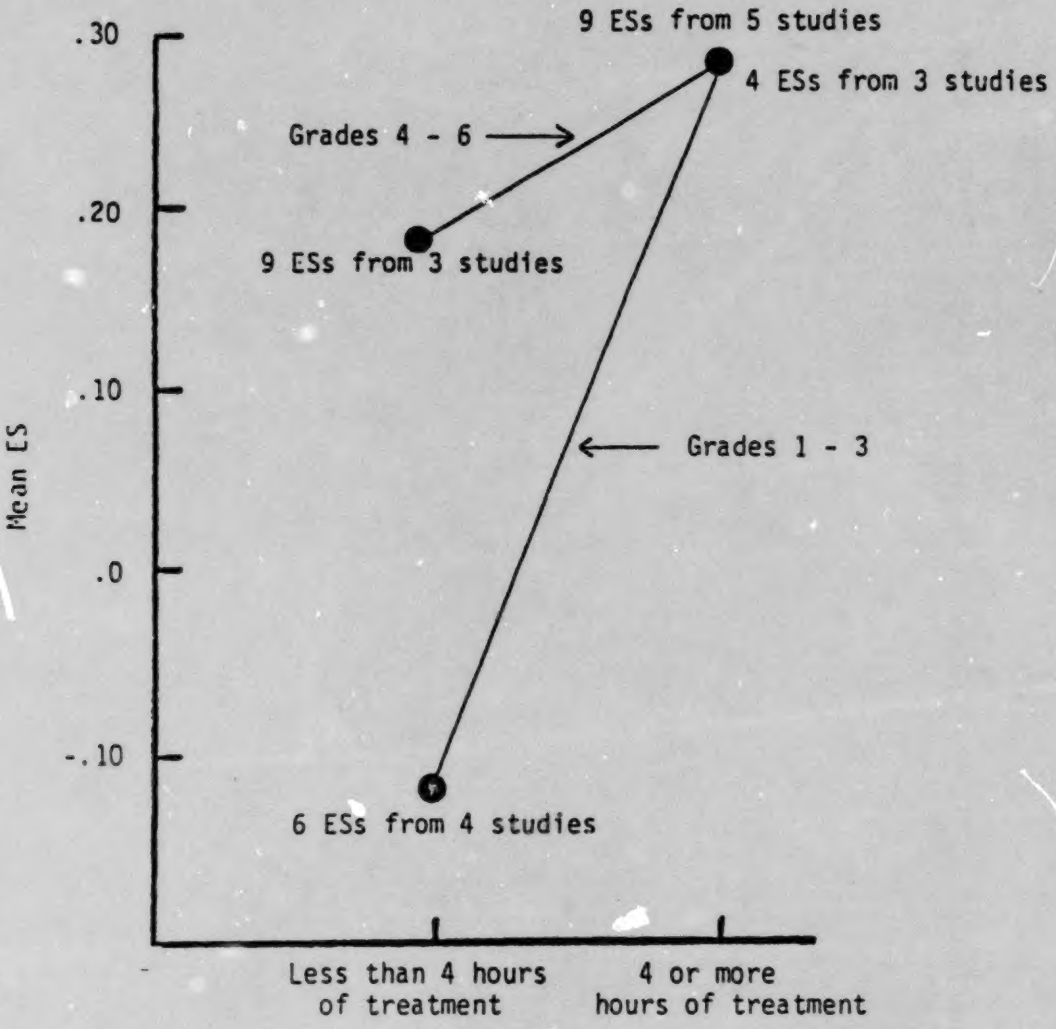
*Achievement test scores, studies with adequate validity only.

Figure Captions

Figure 1. Mean effect size by treatment length and grade level.

Figure 2. Mean effect size by treatment length and SES.


.50
.40
.30
.20
.10
0
-.10
Mean ES
Low SES
(5 ES's from 5 studies)
Not low SES
(8 ES's from 4 studies)
(12 ES's from 6 studies)
(6 ES's from 4 studies)
Less than 4 hours of treatment
4 or more hours of treatment


.30
.20
.10
.0
-.10
Mean ES
9 ESs from 5 studies
4 ESs from 3 studies
Grades 4 - 6
9 ESs from 3 studies
Grades 1 - 3
6 ESs from 4 studies
Less than 4 hours of treatment
4 or more hours of treatment

The Effects of Training in Test-Taking Skills on Test Performance, Attitudes, and On-Task Behavior of Elementary School Children

Thomas E. Scruggs, Karla Bennion, and Joanne Williams
Utah State University

Running head: EFFECTS OF TRAINING

Abstract

Fifty-eight third graders from two elementary school classrooms were assigned at random to test-training and placebo groups. Students in the test-training group received six sessions of test-wiseness training specifically tailored to the Comprehensive Test of Basic Skills. Students in the placebo group received six sessions of creative writing exercises. The effectiveness of this training on achievement tests was obscured due to the presence of ceiling effects. Supplementary analyses, however, provided some support for the effectiveness of this training. Trained and untrained groups were not seen to differ on measures of on-task behavior during the testing situation. An analysis of reported attitudes toward tests taken immediately after the three-day testing period indicated that (a) the standardized test experience was a stressful one for control subjects, and (b) that the test-wiseness training had exerted a significant ameliorating effect in the treatment group. Results indicated that test-wiseness training may reduce levels of anxiety in elementary school children during test situations.

The Effects of Training in Test-Taking Skills on Test Performance, Attitudes, and On-Task Behavior of Elementary School Children

In recent years, the effectiveness of coaching on achievement test performance has been well studied (Sarnacki, 1979, and Fueyo, 1976, for reviews). In a recent meta-analysis, Bangert-Drowns, Kulik, and Kulik (1983) determined that coaching for achievement tests in the elementary grades produced a generally facilitative effect (ES = .29) over all studies reviewed. More recently, Scruggs, Bennion, and White (1984) have argued that although training in test-taking skills does often produce an effect in the elementary school grades, this effect is dependent upon other factors, for example, length of training, age of students, and economic level of the students trained. Although researchers in the area of test-wiseness training have often looked at variables in addition to actual test scores such as performance on test-wiseness tests and self-esteem, they have not addressed the issue of whether or not such training changes in any way the attitudes of elementary school children toward tests. This in itself could be an important finding for, concerning the degree to which school-age children are subjected to testing procedures, it would be helpful to ensure that such tests were not unnecessarily traumatic. In addition, whether or not training in test-taking skills has a facilitative influence on the level of effort the students put into the test situation remains unclear. Such effort may be evaluated by means of the amount of time on-task students put into the standardized achievement test.

The present investigation was intended to address some of these issues by providing training in test-taking skills to a sample of third grade students and assessing, in addition to test performance, reported attitudes towards the test-taking experience and percent of time actually spent on-task during test administration. Although the effects of test-wiseness training have been well-documented in the past, the present investigation was intended to shed some light on peripheral issues and to address more specifically exactly what changes in attention and attitude occur as a result of coaching on achievement tests.

Method

Subjects

Subjects were 58 elementary-age school children attending the third grade in two different classrooms at a western rural school district. Sex was evenly distributed. Subjects were selected at random from both classes to participate in treatment and placebo groups.

Materials

Materials included a manual with six scripted 20- to 30-minute lessons in test-taking skills specifically tailored to the

Comprehensive Test of Basic Skills, Level E. These materials were developed specifically for this project and also included student workbooks for practice activities by the students on the Reading Subtest of this test (Williams, 1984).

Procedure

Over a two-week period, treatment students were taught six lessons in test-taking skills appropriate to the Reading Subtest of the California Test of Basic Skills. These lessons included, for example, time-using strategies, deductive reading strategies, error avoidance strategies, and specific practice activities in each of the subtests. To control for possible Hawthorne effects, the placebo group was given six exercises in creative writing at the same time treatment students were receiving test training. Immediately prior and immediately after training, students in the training group were given pre and posttests of test-taking skills to determine whether students had learned from the training package. This measure is shown in Table 1. Within three days after the conclusion of training, students were given the California Test of Basic Skills by their regular classroom teachers in their regular instructional class. During the taking of this test, observational measures were taken on on-task

---

Insert Table 1 about here

---

behavior of students by four trained observers unaware of group memberships of the students being observed. The observers employed a time-sampling procedure on an interval of 30 seconds. Each student observed was observed for 30 minutes. On-task behavior was computed as percentage of times sampled on-task during actual test performance and on-task behavior while directions were being given. On-task behavior during directions was defined as orientation of student's eyes toward either teacher or test booklet and pencil-and-paper compliance with accompanying sample activities. On-task during testing was defined as student's eyes directed toward test booklet, pencil in hand, activity marking, reading, or asking teacher direct questions with specific reference to the test. After completion of the third and final day of testing, students were given an attitude toward tests questionnaire (see Figure 1). This questionnaire consisted of 10 items in an agree/disagree format. Students completed the questionnaire together while the teacher read items to the class.

Insert Figure 1 about here

Results

Pre and posttest scores of the treatment group on the measure of test-taking skills were completed by means of a correlated t test. On average, students scored 41% correct on the pretest, and 89% on the posttest. These differences were statistically significant ($\underline{t}(27) = 13.9$, $\underline{p} < .001$).

Mean scores on the Reading subtest of the CTBS were computed and compared statistically by means of $\underline{t}$ tests. As can be seen in Table 1, none of the group differences are statistically significant. Interpretation is not possible, however, due to the presence of overwhelming ceiling effects exhibited on all subtests.

Insert Table 1 about here

A supplementary analysis was conducted on the lower half of each group chosen by the previous year's total reading scores and is given in Table 2. This analysis indicates that standardized gain scores between second and third grade testing were significantly higher in favor of the treatment group on Word Attack Subtest and Total Reading Score.

Insert Table 2 about here

On-Task Behavior

Mean on-task behavior during directions, during testing, and total is given in Table 1. As can be seen, no significant group differences were found.

Attitudes Toward Tests

Reliability of the attitude measure was computed by means of a Kuder-Richardson 20 formula and was given at .88, indicating a moderately strong degree of internal consistency for a measure of this type. Differences between the mean scores of the two groups were nonsignificant, t less than 1 in absolute value. An inspection of Figure 2, however, shows that the distribution of these two groups differs strongly. These figures are most obvious when one employs a curve-smoothing technique of combining the mean scores for each of two adjacent frequencies and are given in the same figure. The difference between these dispersions was tested statistically in two ways: mean differences from the mean in standard scores were computed for subjects in each group and compared statistically. The mean distance from the mean of the placebo group was statistically greater than the average distance from the mean in the training group ($p < .01$). In addition, a Kolmogorov-Smirnov two-sample test was applied to each half of the distribution. For the lower half of each distribution (that is, students scoring 0 through 5 on the measure), the distributions were statistically different ($Z = 1.529$, $p < .02$), while the upper half of each distribution was not seen to differ significantly ($Z = .756$, $p = .617$).

Insert Figure 2 about here

Discussion

The present investigation does not offer conclusive evidence that the particular training package employed significantly improved test scores, due to the ceiling effects reported in the Results section. However, it is thought that many students did benefit from this training for the following reasons: (a) students in the lower half of the treatment group exhibited statistically higher gain scores over the previous year's testing than did the lower half of the placebo group, (b) students in the treatment group scored significantly higher on a posttest of test-taking skills than they had on the pretest before training, and (c) reviews of many studies previously conducted (see Scruggs, Bennion, & White, 1984) indicate that this type of training generally has facilitative effects on test-taking performance. Particularly, this training previously demonstrated a significant effect on a subtest similar to the Word Attack subtest in a sample of learning disabled and behaviorally disordered children (Scruggs, 1984).

That achievement test coaching results in greater levels of on-task behavior on the part of students was not supported by the present investigation. Student on-task behaviors while listening to directions and while taking the test itself were very similar.

Analysis of the attitude data did suggest that students in the treatment group reported more "normal" attitudes than those in the placebo group. The abnormal distribution of scores in the placebo group is highly reminiscent of that of a population under stress (see Wilson, 1973). The fact that the abnormally high number of very negative attitudes was not present in the treatment condition while the number of strongly positive attitudes was relatively similar suggests that this treatment may have contributed to more positive attitudes on the part of those students who may otherwise have developed strong negative reactions to the test and the test-taking situation. It should be noted here that completely positive attitudes toward tests was not expected and is not necessarily a realistic expectation. What was expected was a roughly normal distribution centering around the mean of about 5, which is in fact the distribution seen in the training group. The large proportion of extreme scores in the placebo group (with fully two-thirds of the scores within 1 point of 0 or 10) indicates that the population had been subjected to some stress and had reported widely polarized views on the test-taking process. In the training group, these attitudes seemed to have been ameliorated substantially.

References

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test scores. Review of Educational Research, 53, 571-585.

Fueyo, V. (1977). Training test-taking skills: A critical analysis. Psychology in the Schools, 14, 180-184.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 49, 252-279.

Scruggs, T. E., Bennion, K., & White, K. R. (1984). Improving achievement test scores in the elementary grades by coaching: A meta-analysis. Unpublished manuscript, Utah State University.

Williams, J. (1984). Super score: Training materials for the CTBS. Unpublished materials, Utah State University.

Wilson, G. D. (1973). The concept of conservatism. In G. D. Wilson (Ed.), The psychology of conservatism. New York: Academic Press.

Footnote

Preparation of this manuscript was supported in part by Department of Education Grant #84.023, Research in the Education of the Handicapped. The authors would like to thank Clyde Bartlett, Principal, and Loila Anderson and Edna Eams, teachers, at Wilson Elementary School, Logan, Utah. We would also like to thank Marilyn Tinnakul for her assistance in the preparation of this manuscript. Address requests for reprints to: Thomas E. Scruggs, Ph.D., UMC 68, Utah State University, Logan, Utah 84322.

Table 1

T-Tests by Group

CTBS Reading Subtests

Variable	N	$\bar{X}$	SD	T	2-tail prob.
Word attack					
Tx	29	29.79	4.87	.05	.959
Cx	29	29.72	5.37		
Vocabulary					
Tx	29	26.31	4.58	-.49	.624
Cx	29	26.90	4.47		
Comprehension					
Tx	29	26.48	4.06	.79	.434
Cx	29	25.51	5.21		
Total reading					
Tx	29	82.59	12.35	.13	.898
Cx	29	82.14	14.04		

Table 1 (continued)

Variable	N	$\bar{X}$	SD	T	2-tail prob.
CTBS total battery					
Tx	29	150.17	24.68	-.60	.549
Cx	29	154.03	24.10		
Attitude toward test-taking					
Tx	29	5.59	2.97	.59	.557
Cx	27	5.04	3.95		
On-task during directions					
Tx	18	45.28	15.78	-.75	.458
Cx	18	50.06	21.89		
On-task during testing					
Tx	18	77.67	16.18	.07	.941
Cx	18	77.28	14.98		
Total on-task					
Tx	18	65.78	14.76	-.45	.656
Cx	18	67.78	11.82		

Table 2

Gain Score Differences Between the Lower Half of Each Group (Chosen by Last Year's Total Reading)

Variable	*N*	$\bar{X}$	*SD*	Error	*T*	Prob.
Word attack						
Tx	12	25.83	39.55	11.42	2.41	.012
Cx	14	-20.86	47.06	12.58		
Vocabulary						
Tx	12	18.67	50.77	14.66	.49	.625
Cx	14	7.93	58.69	15.69		
Comprehension						
Tx	12	53.17	37.96	10.96	1.46	.158
Cx	14	24.79	57.54	15.38		
Total of all subtests						
Tx	12	97.67	52.64	15.20	2.51	.019
Cx	14	11.86	107.92	28.84		

Figure Captions

Figure 1. Attitude measure.

Figure 2. Distribution of attitude scores.

Circle YES or NO.

YES NO 1. Taking a test is my favorite thing to do at school.

YES NO 2. Sometimes I am nervous when I take a test.

YES NO 3. I look forward to taking a test.

YES NO 4. I dislike taking a test when I don't know the answers.

YES NO 5. I wish we had fewer tests.

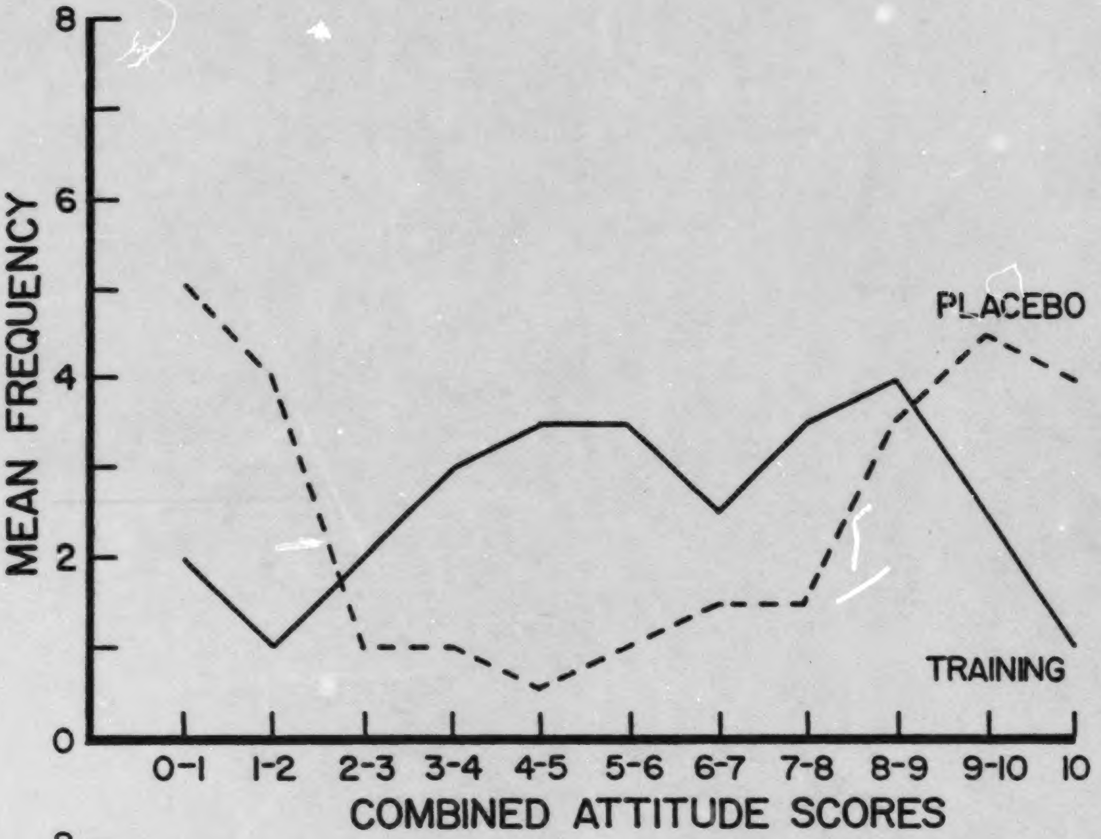
YES NO 6. Taking a test is always fun.

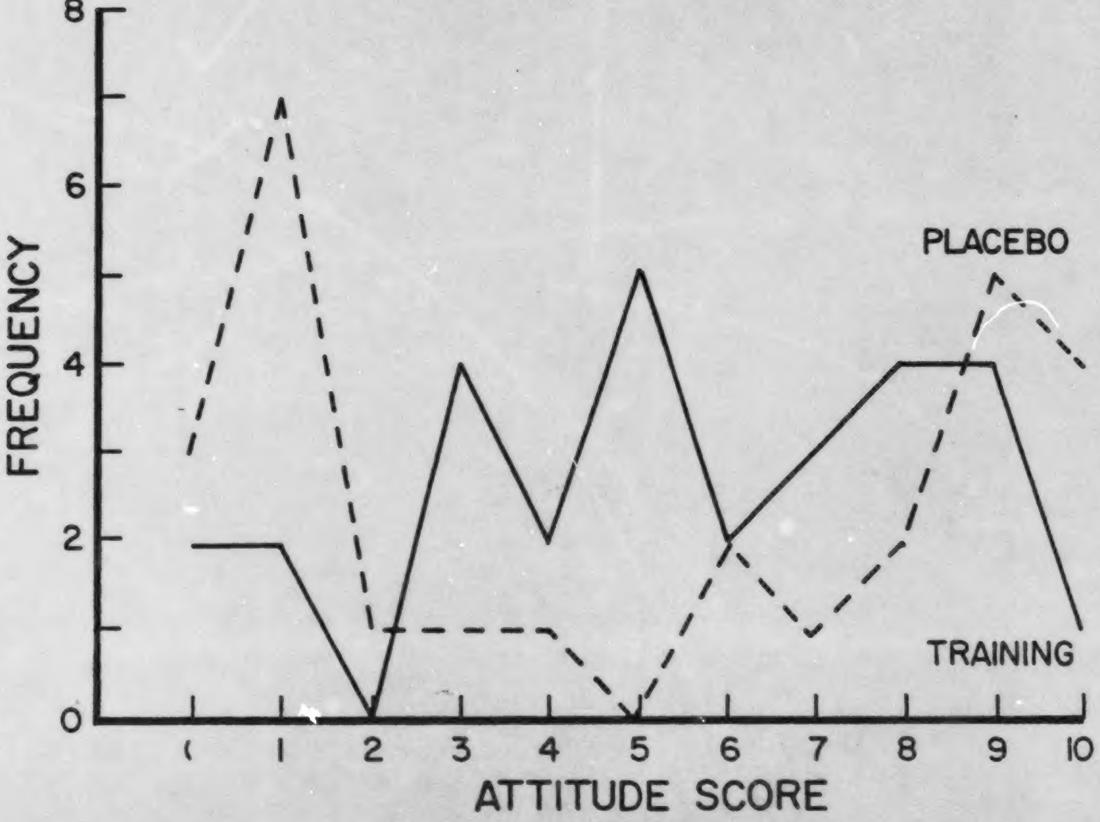
YES NO 7. I like tests even when I don't know the answers.

YES NO 8. Taking a test is one of the worst things about school.

YES NO 9. I would rather do something else besides take a test.

YES NO 10. I wish we had more tests.


MEAN FREQUENCY
0
2
4
6
8
PLACEBO
TRAINING
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
9-10
10
COMBINED ATTITUDE SCORES


FREQUENCY
0
2
4
6
8
PLACEBO
TRAINING
0
1
2
3
4
5
6
7
8
9
10
ATTITUDE SCORE

Teaching Test-Taking Skills to Learning Disabled and Behaviorally Disordered Children

Thomas E. Scruggs

Utah State University

Running head: TRAINING TEST-TAKING SKILLS

Abstract

Ninety-two second, third, and fourth grade children classified as learning disabled (LD) or behaviorally disordered (BD) were randomly assigned to treatment and control groups. Students assigned to the treatment condition were taught test-taking skills pertinent to reading achievement tests. Students were taught in small groups over a two-week period in such strategies as attending to appropriate stimuli, marking answers carefully, time using, and error avoidance. Following the training procedures, students were administered standardized achievement tests in their regular classroom assignments. Results indicated that third and fourth grades scored significantly higher on the word study skills subtest and descriptively higher in other subtests of the Stanford Achievement Test. Second grade students did not appear to have been affected by the training. The relevance of the training of this test to other tests involving problem-solving strategies is discussed.

Teaching Test-Taking Skills to Learning Disabled and Behaviorally Disordered Children

Successful performance in school is to a great extent dependent upon the application of effective learning and problem-solving strategies on academic tasks. Students are often called upon to meet particular format and task demands on academic assignments with effective strategies for dealing with these tasks and successfully completing them. Much of the failure of learning disabled (LD) students in school-related tasks has been attributed to a lack of ability in applying such problem-solving strategies (Reid & Hresko, 1980). A body of literature has been established in recent years which documents difficulties of learning disabled students in employing appropriate learning and problem-solving strategies in school. Particular deficits have been noted in the areas of: (a) attending to the critical components of a task (Atkinson & Seuneth, 1973; Hallahan & Reeve, 1980; Hallahan, Kauffman, & Ball, 1973; Ross, 1976; Tarver, Hallahan, Kauffman, & Ball, 1976), (b) selecting a strategy appropriate to addressing a particular academic task (Mastropieri, Scruggs, & Levin, in press; Torgesen, 1977; Torgesen & Goldman, 1977), and (c) effectively employing appropriate problem-solving strategies (Hallahan, 1975; Spring & Capps, 1974; Torgeson, Murphy, & Ivey, 1979).

Given the above documented deficiencies, it would appear that one area of particular difficulty for learning disabled and perhaps other mildly handicapped children would be the problem-solving strategies necessary for successful completion of standardized achievement tests. These group-administered tests typically expect learners to function individually in large-group situations, effectively employ time constraints, and develop and employ strategies specifically suited to answering questions which may be ambiguous or to which the answers are often not completely known (Haney & Scott, 1980). Some recent research with learning disabled students indicates that these students do, in fact, exhibit deficiencies with respect to use of effective strategies in standardized test-taking situations. Scruggs and Lifson (1984) administered questions from standardized reading comprehension tests to LD and non-LD students without providing the accompanying reading passages. Their results indicated that, although non-LD students were able to answer most "reading comprehension" questions without reading the accompanying passages, LD students were not able to do this. This investigation reiterated previously made questions concerning what reading comprehension tests actually measure, and also suggested that many LD students may lack some specific test-taking strategies, such as ability to effectively employ partial and/or prior knowledge. Drawing upon a previous investigation with mostly nondisabled children (Scruggs, Bennion, & Lifson, in press), Scruggs, Bennion, and Lifson (1984) recently interviewed learning disabled children with respect to the manner in which they had interpreted and answered reading achievement test items. Analysis of this strategy reports indicated that (a) LD students were less likely to select and utilize strategies appropriate to different types of test questions, and (b) LD students were more likely to be negatively influenced by misleading distractors. Such results suggested that learning disabled and perhaps other mildly handicapped populations may have more difficulty than other students adapting to specific task and format demands of standardized achievement tests and, consequently, resulting scores may be less valid estimations of potential performance than those of other students. Although any observed deficit in "test-taking strategies" on the part of learning disabled children would be expected to be representative of more global problem-solving strategy deficits in school-related tasks on the whole, it may be possible that specific training in test-taking skills may be particularly beneficial to children referred for mild learning and/or behavior problems. Many attempts have been previously made to improve achievement test scores by coaching in test-taking skills, but the results have been somewhat mixed and have appeared to affect different populations differentially. For example, Scruggs, Bennion, and White (1984) in a recent meta-analysis reported that students from the lower grade levels and students from low economic backgrounds tended to differentially benefit from extended training in test-taking skills. This finding, although not directly relevant to special education, does imply that these students may be taught some of the critical skills that they apparently lack when confronted with standardized achievement tests. It was the purpose of this investigation to determine whether such skills could be taught and whether such skills could, in fact, increase performance on standardized achievement tests without an accompanying increase in knowledge of the task being assessed.[1]

Method

Subjects

Subjects were 92 second, third, and fourth grade students attending resource rooms or self-contained classes in a large western school district. Twenty-five students were second graders, 37 were third graders, and 31 were attending fourth grade classes. The 68 boys and 34 girls had tested at an average of the 20th percentile (*SD* = 9.3) at the previous year's testing in reading. Thirty-nine students were classified as LD, and 54 students were classified as behaviorally disordered according to Public Law 94-142 and local school district criteria (for learning disabilities, this included a 40% discrepancy between ability and achievement). Twenty-two students were enrolled in self-contained classes, and 70 students were attending resource rooms.

Materials

Materials were developed as part of a larger project involving improving test-taking skills of LD and BD elementary students (Taylor & Scruggs, 1983) and consisted of eight scripted lessons for each grade level in a direct instruction format and accompanying workbooks for students which included pencil-and-paper practice activities (Scruggs & Williams, 1984). The general test-taking strategies taught in these materials included attending, marking answers carefully, choosing the best answer carefully, error avoidance strategies, and appropriate situations for soliciting teacher attention. In addition, specific test-taking strategies were taught for each specific reading subtest relevant to reading in the Stanford Achievement Test. These included structured practice in specific test formats for each subtest and specific application of general test-taking strategies to each specific subtest. For example, with respect to the letter-sound subtest, students were taught to employ the following sequence of strategies:

1. Look at the first word; read it.

2. Pronounce to yourself and think of the sound of the underlined letter.

3. Carefully look at the answer choices and choose the word with the same sound as the underlined letter.

4. If you don't know all the words, read the words you do know, or read parts of individual words that you may know.

5. If you are not sure of the answer, see if there are some answers that you are sure are not correct, and eliminate those.

6. Color in the answer quick, dark, and inside the line.
7. Never skip an answer.

Procedure

Experimental subjects were taught in small groups ranging from one to five in size and were taught four 20-minute lessons per week, for two weeks. Positive responding and attention to task was reinforced with stickers. Immediately prior to the training sessions, and immediately after the last training session, students were administered a criterion test of the skills which were taught (see Figure 1). This test was a 10-item test of test-taking skills including questions about time using, question asking, and elimination strategies. The first seven sessions taught the use of test-taking strategies within the specific context of each of the reading-related subtests. The last session consisted of a general review of all previous procedures. Each day of instruction involved extensive work with practice activities applied to practice test items. At no time during this training procedure were subjects taught any information concerning the content of the test which was not given in the published test directions. Within five days of the training procedure, students were administered as a group the Stanford Achievement

Insert Figure 1 about here

Test. This administration was done in the regular or self-contained classroom settings by their regularly assigned teacher. Although teachers were aware of the membership of each student in the experimental group, response protocols were scored by machine.

Results

Pre and posttests of the experimental students on the criterion measure were compared statistically by means of a correlated t test. It was found that the performance on the posttest was significantly higher than pretest scores ($p < .01$). Students scored an average of 40% percent correct on the pretest, and 77% correct on the posttest.

Summary of analyses are given in Table 1. Data for second grade students were analyzed separately because (a) sufficient test data from previous years' testing existed to compute analysis of covariance, and (b) patterns of effects of treatment appeared to be somewhat different in this group. Although second grade subjects were assigned at random to experimental and control groups, they differed significantly ($p < .05$) with respect to previous years' testing, and, therefore, analyses must be interpreted with caution. Although raw scores on reading subtests

Insert Table 1 about here

in fact favored the control group, these differences were decreased substantially by the use of previous years' testing as a covariate. In spite of this adjustment with the covariate, the second grade control group apparently statistically outperformed the treatment group in the comprehension subtest. Since the groups did differ significantly in the year's previous testing, however, and since a similar comprehension subtest was not a part of the first grade test battery, which likely weakened the covariate, this finding appeared to be an artifact of selection bias. Third and fourth grade data were also analyzed separately. However, since in the third and fourth grade students, over one-third of the total sample were missing previous years' test scores, it was not possible to use previous years' testing as a covariate. As can be seen in Table 1, differences generally favored treatment groups although none of the initial findings were significant to the .05 level. However, the treatment effect was replicated over third and fourth grade groups with a particular effect seen in the Word Study subtest raw scores. Effect sizes were .63 in the third grade students, and .48 with the fourth grade students, both in favor of the treatment group. An evaluation of third and fourth grade combined using scale scores, however, indicates a significant treatment effect for the experimental students on the Word Study Skills subtest. Although comprehension scores and total reading scores also favor the treatment group, these differences are not statistically significant.

Discussion

The analysis of pre and posttest scores indicated that test-taking skills could be successfully taught to this sample of second, third, and fourth grade learning disabled and behaviorally disordered children. The fact that significant gains were made in these critical skills indicates that learning disabled and behaviorally disordered children at this age level do, in fact, lack certain test-taking skills which are potentially helpful in taking standardized achievement tests.

An analysis of the data apparently suggests that second grade students did not benefit from the training package. These data are difficult to interpret accurately, however, considering the fact that this group of children had scored significantly lower than the control group on tests administered one year previous. Although the use of analysis of covariance somewhat compensated for these differences, any interpretation of the results must be made with caution considering such significant differences existed between the two groups in the first place. However, considering these were reading tests and that the average reading performance of second grade learning disabled and behaviorally disordered children is extremely low, it may be that second grade special education students lack sufficient reading skills in order to make the most of training in test-taking skills. This may indicate that it is more prudent to wait until certain critical reading skills have been mastered before training of this nature will be beneficial. Considering the previous differences between the experimental and control group with respect to the second grade population, however, this interpretation cannot be made conclusively. Analysis of the third and fourth grade data indicated that training in test-taking skills did significantly increase scores on the Word Study Skills subtest of the Stanford Achievement Test for third and fourth grade learning disabled and behaviorally disordered students. Differences favoring the treatment group were also found in all the subtests and total reading score, although these differences were not significant. The fact that the Word Study Skills subtest was increased significantly may be a function of the fact that this particular subtest involves many format changes over a short period of time and thus, was more amenable to increased performance through guided practice and feedback on successful skills necessary for completion of the subtest (Bennion, Scruggs, & Lifson, 1984). Since previous research has indicated that learning disabled children are more likely to have difficulty with formats on this type of subtest (Scruggs, Bennion, & Lifson, 1984), this seems a likely explanation for the fact that Word Study Skills performance was significantly facilitated. The degree of facilitation of this subtest in scale score points apparently compares to a gain of three academic months for the average student receiving this treatment. This gain is consistent with the findings of a recent meta-analysis (Scruggs, Bennion, & White, 1984) which indicated that other students tended to gain approximately two to three months in situations involving extended training on test-taking skills. Although a three-month gain does not seem particularly large, it must be weighed against the finding that this was accomplished in eight relatively short lessons over a two-week period and that training in reading skills over the same period would be unlikely to produce such a gain. However, any gain at all which is not the result of training in the associated content areas indicates the possibility that some of the error variance in this test is being eliminated and, in fact, Table 1 indicates descriptively that standard deviations were consistently lower in treatment groups than control groups. This finding is not conclusive but does suggest that error was reduced on the part of treatment children.

Overall, the findings indicate that critical test-taking skills can be taught to learning disabled and behaviorally disordered second, third, and fourth grade children and that these skills tend to raise these students' performance on standardized achievement tests.

References

Atkinson, B. R., & Seunath, O. H. M. (1973). The effect of stimulus change in attending behavior in normal children and children with learning disorders. Journal of Learning Disabilities, 6, 569-573.

Hallahan, D. P. (1975). Comparative research studies on the psychological characteristics of learning disabled children. In W. M. Cruickshank & D. P. Hallahan (Eds.), Perceptual and learning disabilities in children, Vol. 1. Psychoeducational practices. Syracuse, NY: Syracuse University Press.

Hallahan, D. P., & Reeve, R. E. (1980). Selective attention and distractibility. In B. Keogh (Ed.), Advances in special education (Vol. 1). Greenwich, CT: Jai Press.

Hallahan, D. P., Kauffman, J. M., & Ball, D. W. (1973). Selective attention and cognitive tempo of low achieving and high achieving sixth grade males. Perceptual and Motor Skills, 36, 579-583.

Haney, W., & Scott, L. (1980). Talking with children about tests: A pilot study of test item ambiguity. National Consortium of Testing Staff Circular No. 7. Cambridge, MA: The Huron Institute.

Mastropieri, M. A., Scruggs, T. E., & Levin, J. R. (in press). Memory strategy instruction with learning disabled adolescents. Journal of Learning Disabilities.

Reid, D. K., & Hresko, W. P. (1980). Thinking about thinking about it in that way: Test data and instruction. Exceptional Education Quarterly, 1(3), 47-57.

Ross, A. O. (1976). Psychological aspects of learning disabilities and reading disorders. Nwe York: McGraw-Hill.

Scruggs, T. E., Bennion, K., & Lifson, S. (in press). An analysis of children's strategy use on reading achievement tests. Elementary School Journal.

Scruggs, T. E., Bennion, K., & Lifson, S. (1984). Spontaneously employed test-taking strategies of high and low comprehending elementary school children. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Scruggs, T. E., Bennion, K., & White, K. (1984). Improving achievement test scores in the elementary grades by coaching: A meta-analysis. Unpublished manuscript, Utah State University.

Scruggs, T. E., & Lifson, S. (1984). Are learning disabled students "test-wise?": An inquiry into reading comprehension test items. Unpublished manuscript, Utah State University.

Scruggs, T. E., & Williams, J. (1984). Super score. Unpublished training materials, Utah State University.

Spring, C., & Capps, C. (1974). Encoding speed, rehearsal, and probed recall of dyslexic boys. Journal of Educational Psychology, 66, 780-786.

Tarver, S. G., Hallahan, D. P., Kauffman, J. M., & Ball, D. W. (1976). Verbal rehearsal and selective attention in children with learning disabilities: A developmental lag. Journal of Experimental Child Psychology, 22, 375-385.

Taylor, C., & Scruggs, T. E. (1983). Research in progress: Improving the test-taking skills of learning disabled and behaviorally disordered elementary school children. Exceptional Children, 50, 277.

Torgesen, J. K. (1977). The role of nonspecific factors in the task performance of learning disabled children: A theoretical assessment. Journal of Learning Disabilities, 10, 27-34.

Torgesen, J. K., & Goldman, T. (1977). Verbal rehearsal and short-term memory in reading-disabled children. Child Development, 48, 56-60.

Torgesen, J. K., Murphy, H. A., & Ivey, C. (1979). The influence of an orienting task on the memory performance of children with reading problems. Journal of Learning Disabilities, 12, 396-401.

Footnote

The preparation of this manuscript was supported in part by a grant from the Department of Education, Special Education Programs. The authors would like to thank Dr. Joyce Barnes and the teachers and administrators of the Granite School District for their cooperation and assistance. The authors would also like to thank Marilyn Tinnakul for her assistance in the preparation of this manuscript.

[1] The usefulness of standardized achievement tests in special education has been, and remains, a controversial issue (Salvia & Ysseldyke, 1979), which is not intended to be addressed by the results of the present investigation. This investigation was undertaken to determine whether the problem-solving strategies of the type needed for the successful completion of achievement tests could be trained. An additional assumption was that reduction of possible measurement error, on any assessment instrument in common use, is desirable.

Table 1

Test Score Data

2nd Grade - Analysis of Covariance

Variable	N	$\bar{X}$	SD	Adj. Mean	F	Prob.
Word reading						
Tx	12	15.58	4.32	17.00	1.01	.326
Cx	13	20.77	7.65	19.41		
Comprehension						
Tx	12	16.42	6.35	18.51	5.10	.035
Cx	13	26.18	9.00	24.08		
Word study						
Tx	12	25.67	5.69	29.44	.47	.50
Cx	13	31.62	10.05	27.49		
Total reading						
Tx	12	57.67	14.34	63.01	2.58	.124
Cx	13	78.38	22.60	72.93		

Table 1 (continued)

3rd Grade

Variable	*N*	$\bar{X}$	*SD*	*T*	2-tail prob.
Comprehension raw scores					
Tx	18	24.61	7.59	-.36	.725
Cx	19	25.79	11.98		
Word study raw scores					
Tx	17	29.12	8.09	1.70	.099
Cx	19	24.95	6.65		
Total reading raw scores					
Tx	18	52.06	16.21	.24	.813
Cx	19	50.74	17.33		
Total battery scaled scores					
Tx	17	564.00	17.80	.00	.999
Cx	19	564.00	21.09		

Table 1 (continued)

4th Grade

Variable	N	$\overline{X}$	SD	T	2-tail prob.
Comprehension raw scores					
Tx	17	17.71	7.50	.61	.545
Cx	14	15.79	9.96		
Word study raw scores					
Tx	17	26.53	10.12	1.28	.209
Cx	14	21.93	9.68		
Total reading raw scores					
Tx	17	44.24	16.54	1.05	.303
Cx	14	37.71	18.02		
Total battery scaled scores					
Tx	17	572.35	26.15	.04	.968
Cx	14	572.90	20.60		

Table 1 (continued)

3rd and 4th Grades Combined

Variable	N	X̄	SD	Standard error	T	df	2-tail prob.
Comprehension scaled scores							
Tx	35	559.00	30.58	5.17	.41	65	.680
Cx	32	556.00	38.77	6.85			
Word study scaled scores							
Tx	34	578.00	31.66	5.43	2.26	65	.027*
Cx	33	562.00	28.04	4.88			
Total battery scaled scores							
Tx	34	568.00	22.43	3.85	.15	65	.883
Cx	33	567.00	20.95	3.65			

Figure Caption

Figure 1. Pre-post test.

1. When I don't understand the teacher,
   ○ I go up to the teacher.
   ○ I raise my hand.
   ○ I ask another student.

2. When I mark outside the answer bubble,
   ○ I mark it carefully.
   ○ I can not erase and fix it.
   ○ I might get the answer wrong.

3. After I read the test question,
   ○ I read all the answer choices.
   ○ I think and choose the best answer.
   ○ I guess the best answer.

4. A vocabulary test asks
   ○ the meaning of a word.
   ○ how to read a word.
   ○ how to spell a word.

5. The stop sign tells me to
   ○ stop and then go on.
   ○ stop and check my work.
   ○ stop and lay my pencil down.

6. When I can't read all the words in the answer choices,
   ○ I read the words I know first.
   ○ I guess the answer first.
   ○ I go on to the next question.

7. When I don't know the answer,
   ○ I skip the question.
   ○ I guess the best answer.
   ○ I raise my hand.

8. When I take a comprehension test,
   ○ I read the answer choices first.
   ○ I read the questions first.
   ○ I read the passage first.

9. When I take a syllables test, I look
   ○ for a compound word.
   ○ for a word that has a prefix
   ○ for a word that is divided the right way.

10. The letter-sound in a letter-sounds test
   ○ can be spelled by different letters.
   ○ are always in the middle of the word.
   ○ are always spelled with the same letters.